Washington University in St. Louis
Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations

Arts & Sciences

Spring 5-15-2019

Exploring Infant Leukemia through Exome Sequencing and an In Vitro Model of Hematopoietic Development

Mark Cannon Valentine
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

 Part of the Developmental Biology Commons, Genetics Commons, and the Oncology Commons

Recommended Citation

Valentine, Mark Cannon, "Exploring Infant Leukemia through Exome Sequencing and an In Vitro Model of Hematopoietic Development" (2019). *Arts & Sciences Electronic Theses and Dissertations*. 1798.
https://openscholarship.wustl.edu/art_sci_etds/1798

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Human and Statistical Genetics

Dissertation Examination Committee:
Todd Druley, Chair
Grant Challen
Mary Dinauer
Jeffrey Magee
Christopher Sturgeon
David Wilson

Exploring Infant Leukemia through Exome Sequencing and an *In Vitro* Model of Hematopoietic Development

by

Mark Cannon Valentine

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019

St. Louis, Missouri

© 2019, Mark Cannon Valentine

Table of Contents

List of Figures

List of Tables

Acknowledgments

This work was only accomplished with the help and support of many individuals who contributed, either directly to the scientific aspects of this dissertation, or provided emotional, personal or mental health support over the years.

Dr. Todd Druley provided ideas, feedback, samples, guidance, encouragement, funding and a unique lab environment conducive to creative and independent work. His willingness to allow me to follow this project wherever it led, regardless of our familiarity with the new territory is laudable and has resulted in a much more complete and interesting story than would have been possible with a more conservative approach. I am grateful to him for the time I spent in his lab, and for the learning and growth that occurred under his supervision.

Dr. Chris Sturgeon was an unofficial co-thesis mentor. He taught me much about developmental hematopoiesis specifically, but also about how to best work and behave as a research scientist generally. His mentorship, knowledge and example have been invaluable resources in my training.

The other members of my thesis committee, Grant Challen, Dave Wilson, Mary Dinauer and Jeff Magee have provided guidance and feedback that have critically and meaningfully shaped this project.

Andrew Young and I began our journey as MSTPs together in 2010, and continued as new graduate students in the Druley lab in 2012. Along the way we have roasted pigs, eaten more sushi than any reasonable human should, and all the while discussed science, cancer, medicine, life, relationships and everything in between. He has been a support and friend who made my meandering journey richer and more enjoyable.

The Druley and Sturgeon lab members have been my collaborators, friends, drinking buddies and de facto therapists. Drew Hughes, Katie Thornton, Sara Chasnoff, Spencer “Spatula” Tong, Wing Hong, Maggie Ferris, Carissa Dege, Phillip Creamer and Kendra Sturgeon in particular have been meaningful players in my graduate school experience.

I have learned and benefitted from the knowledge insights and expertise of many other members of the Washington University community. I thank my fellow trainees, their PIs and the support staff with whom I have interacted over the years.

I feel privileged and honored to be a member of the MSTP entering class of 2010. My fellow classmates are an inspiring and promising group. I hope to live up to the standards that they achieve and look forward to watching their successes and contributions to humanity.

I thank my parents for their guidance, encouragement and support through the years, and my siblings for their support and examples. I am particularly grateful to my son, Dillon, who brings joy and happiness to everyone he meets. His goodness and gentleness makes me always strive to be better and restores hope for the future. His patience during the many, many hours that he spent in lab with me instead of swimming, camping, fishing, wresting or hiking was more than any parent could expect. I look forward to spending time doing more of these activities with him in the future.

This work was funded by the MSTP training grant as well as the Genome Analysis Training Program (GATP)

Mark Valentine, Washington University in St. Louis, May 2019

Dedicated to DFV.

Abstract of the Dissertation

Exploring Infant Leukemia through Exome Sequencing and an *In Vitro* Model of Hematopoietic Development

by

Mark Cannon Valentine

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2019

Todd Druley, Chair

Cancer is a heterogeneous disease with myriad causes and outcomes. Many of the cancers that occur in adult populations have become increasingly well characterized with the advent of affordable high-throughput sequencing. These studies have revealed that cancer is largely a disease of somatic mutation in the adult population. In strong contrast to this, childhood cancers have an exceedingly low rate of somatic mutation. At the extreme end of this spectrum is Infant Leukemia (IL). Sequencing of IL has revealed that these tumors generally have one or fewer somatic SNP. In the absence of a somatic explanation for IL, many other possible explanations have been put forth. To date, however, none of these has been able to fully explain the incidence of this disease. In this context, we hypothesized that inherited germline variation, rather than somatically acquired mutations, played a role in the development of IL. We showed that IL patients have an excess of rare, non-synonymous, inherited variation in known-leukemia associated genes. We further showed that there are several genes that harbor far more putatively damaging variation in IL patients than either control exomes, or population databases. These highly variant genes are intolerant of loss-of-function changes, and most perform one of three

critical cellular functions. Together these data suggest that IL is indeed a result of predisposing genetic variation.

Obtaining a clearer understanding of IL has been hindered by the lack of an appropriate model. The fact that this disease arises *in utero* makes it difficult to study *in vivo*, and no animal models have been able to recapitulate the rapid onset of disease. In recent years, several groups have developed *in vitro* models of human hematopoiesis. While these are not yet able to fully capture all aspects of hematopoietic development, they do provide a system in which we can explore the effects of the genetic variation observed in IL patients in a controlled and developmentally relevant setting. Importantly, we are able to effectively separate the primitive and definitive hematopoietic programs and explore each independently, a necessary feature for any IL model. In this work, we present the first steps in the development of a model of IL that is consistent with our sequencing findings. While we do not achieve leukemic transformation, we do show that cells deficient in MLL3, a gene that was frequently variant in our IL cohorts, have a marked impairment in both primitive and definitive hematopoiesis. We find that this is evident both based on surface markers and colony forming ability. In addition to these functional characteristics, we show that the transcriptional and epigenetic profiles of the MLL3-knockout cells are greatly perturbed, consistent with the role of MLL3 as a transcriptional enhancer and epigenetic regulator. These results provide insight into the etiology of IL as a disease of aberrant development, and provide a basis for the establishment of an in vitro model of IL.

1. Introduction

1.1 Infant Leukemia

1.1.1 Cancer

Cancer is a disease that has been affecting humanity for thousands of years[1]. Descriptions of the disease and its causes and treatments are present in ancient Egyptian, Sumerian, Hebrew, Indian, Persian, Chinese and Greek texts[1,2]. The struggle to understand and better treat this mystifying and frequently fatal disease may have begun with these early physicians, but these efforts have grown throughout the years and are continued in earnest today[3–7]. Many breakthroughs have been made, and our understanding of the features, risk factors and driving forces behind cancer is more complete than ever. One of the most important discoveries has been the recognition that various cancers, while linked through some similarities, are actually a diverse group of diseases[2,8–10]. The nature and behavior of a given cancer will be influenced by the tissue and cell of origin, the specific genetic and epigenetic lesions present in the tumor, the age at diagnosis and potentially many other factors[2]. Indeed, cancers arising in different tissues, but containing the same genetic lesion may respond more similarly to certain treatments than cancers from the same tissue, but with different underlying genetics[11–14]. Similarly, a cancer from a given tissue diagnosed in childhood can differ in genetics, treatment susceptibility, and prognosis than one from the same tissue but diagnosed in adulthood[15,16].

This heterogeneity in cancer is further evidenced by the large range in the mutational burden present in different cancer types[17,18]. As expected, cancers with known mutagenic exposures (eg. melanoma, lung adenocarcinoma in smokers) have a much higher rate of mutation than those

without. At the opposite extreme this mutational spectrum are the pediatric cancers, which, as a group, tend to have fewer somatic mutations than their adult counterparts[17,19,20].

While there is clearly a great deal of heterogeneity in cancer, perhaps the strongest link between cancers is that each is a disease of the genome, having a number of genetic changes that leads to the development, growth and spread of cancer[17,21]. Initially, it seems that this link no longer holds in the context of pediatric cancers which, especially in the extreme example of infant leukemia, have far fewer somatic mutations than adult tumors[22]. However, cancer is ever the diverse disease, so it holds that the types of genetic damage that underlie its development would be similarly diverse. Indeed, there are numerous known inherited genetic variants that predispose individuals to cancer[23,24]. While these cancer predisposition genes lessen the requirement for subsequent somatic mutation to develop cancer, they rarely result in cancer in childhood, even less so in infancy. Still, they do provide evidence that inherited germline variants play a role in the development of cancer. It is possible that combinations of inherited variants would lead to more extreme phenotypes. The requirement for the co-occurrence of multiple rare events would potentially explain the rarity of these diseases as well as their extreme early onset. To explore this possibility we turned to the extreme example of infant leukemia.

1.1.2 Features and Outcomes of Infant Leukemia

Infant Leukemia (IL) is defined as any leukemia diagnosed in the first year of life. There are 5 new IL cases diagnosed per 100,000 individuals each year[25]. In contrast to leukemia later in childhood, IL has a dismal 5-year event free survival rate of less than 50%[26,27]. This failure comes despite the achievement of complete remission in more than 90% of patients[26]. Early relapse is common in IL and has not been prevented by either intensified therapy or stem cell transfer[26–28]. There are several negative prognostic factors for IL patients. Younger age at

diagnosis, hyperleukocytosis, poor initial treatment response, immature cell types and the presence of chromosomal translocations involving KMT2A all negatively affect outcomes in these infants[29–31]. Patients who do survive have deficits in organ function, growth and learning throughout later life, presumably due to the intensity of treatment regimens early in life[32–34].

1.1.3 Risk Factors for Infant Leukemia

IL has one of the lowest rates of somatic mutation of any cancer studied to date[22]. This paucity of somatic mutation prompts the question of what is causing IL. Several suggestions that will be outlined below have been made to date, but none adequately explain the incidence of IL. One of the most common features present in IL are chromosomal rearrangements involving KMT2A (also known as MLL) and one of more than 70 fusion partners[30]. IL cases with KMT2A rearrangements (KMT2A-R+) are particularly aggressive. However, KMT2A rearrangements are not necessary for the development of IL, as 20% of IL cases are negative (KMT2A-R-) for this event[27,29]. Also, despite some arguments that MLL-rearrangements alone might drive IL[35] , other studies show that this event alone is insufficient to explain IL[36–38].

Absent somatic events to explain IL, many other possible causes have been explored. A number of environmental factors have been interrogated. In summary, there was no appreciable increase of pediatric cancer incidence as a result of background radiation, non-ionizing radiation, electric fields, childhood infections, vaccinations, breast feeding or daycare attendance[39]. Conversely, maternal consumption of naturally occurring topoisomerase II inhibitors did increase the risk of KMT2A-R+ AML in their infants[40]. Similarly, children with KMT2A-R+ leukemia were more likely to have decreased function in NAD(P)H:quinone oxidoreductase, an association that was even more pronounced in infants[41]. Another study presented evidence that specific variants in the methylenetetrahydrofolate reductase gene conferred protection against the development of

KMT2A-R+ leukemia[42]. Together these studies show that there are environmental risk factors for IL, some of which interact with inherited variants. Even in aggregate, these findings do not fully explain the occurrence of IL, but they do support the notion that inherited variants can play an important role in the development of IL.

1.2 Hematopoietic Development

We hypothesize that inherited variation is responsible, at least in part, for the development of IL. This variation is present from the time of conception. IL also develops *in utero*[43]. This timing suggests that IL might be a result of developmental processes gone awry. In order to explore this idea more fully, we must first understand the normal sequence of events, and the relevant pathways, cell types and processes that occur during normal hematopoietic development. Further, if we wish to model these processes, we will need a tractable system with understood readouts so that we can detect deviations from normal. To achieve these goals, we turn to a human pluripotent stem cell (hPSC) – based directed differentiation system, based on insights into development derived from various model organisms.

1.2.1 Hematopoiesis in the Embryo

Hematopoiesis, the process by which the body forms new blood cells, has been an active area of research for many decades. In the adult human, hematopoiesis occurs in the bone marrow and begins with hematopoietic stem cells, which have the capacity to both self-renew and differentiate into all blood lineages[44]. This adult program is established *in utero*, but is preceded by two waves of earlier hematopoietic programs[45]. The earliest of these, called the primitive program, begins during the third week of gestation[46,47]. Primitive hematopoiesis occurs, not in the embryo proper, but in blood islands located in the yolk sac[48]. Primitive hematopoietic cells

arise concurrently with endothelial cells and are both derived from a common mesodermal precursor, the hemangioblast[49–51]. The primitive program does not give rise to all blood lineages, but is restricted to macrophages, megakaryocytes[47] and primitive erythroblasts, which retain their nuclei and express embryonic globins[52,53]. There are no primitive HSCs, but some of the cells generated by this program might seed developing embryonic tissues and persist there throughout life[54].

A second extra-embryonic wave of hematopoiesis occurs soon after the primitive program. This program generates a population of erythro-myeloid progenitors (EMPs) which are distinct from both the earlier primitive program and the later definitive program[54,55]. The EMPs give rise to an expanded set of lineages relative to primitive progenitors, including definitive erythroid, megakaryocyte, macrophage, neutrophil and mast cell progenitors, but do not have lymphoid potential[55].

The definitive hematopoietic program is also established *in utero*, but arises after the other programs have begun[45]. Like the other programs, definitive hematopoietic progenitors are derived from a subset of endothelial cells that have hematopoietic potential, dubbed hemogenic endothelium[56,57]. The best-studied hemogenic endothelium is the aorta-gonad-mesonephros region, but it is also present in the vitelline and umbilical arteries and the head[54,58,59]. These hemogenic endothelial cells undergo an endothelial-to-hematopoietic transition (EHT) to enter the newly forming circulation[60,61]. These cells represent the first definitive HSCs, and will go on to seed the fetal liver and eventually the bone marrow[45]. The definitive program gives rise to the entire spectrum of hematopoietic lineages: erythroid, myeloid and lymphoid[62].

Decades of research have elucidated many of the genes that are responsible for the various hematopoietic programs. The genes that are required for primitive hematopoiesis frequently encode transcription factors important for many developmental processes. These include LMO2, TAL1 and GATA2[62–69]. Similarly, the definitive hematopoietic program depends on the transcription factors encoded by RUNX1, MYB, CEBPZ, IKZF1, LHX2, REL and SPI1 [70–80]. There are also several genes and gene families that play a role in regulating hematopoiesis that enforce different programs in different species, or act in both the primitive and definitive program. These include the CDX genes, and several HOX genes[81–83]. The induction of these transcription factors depends on input from several signaling pathways including Nodal-activin, bone morphogenetic protein (BMP), WNT-β-Catenin, Fibroblast Growth Factor (FGF), and Notch[84–88]. Clearly, these genes do not form an exhaustive list; new regulators of the complex process of hematopoietic development will continue to be discovered. Still, the knowledge of the various roles that these genes play is critical to both understanding developmental hematopoiesis, and attempting to model this process *in vitro.*

1.2.2 In vitro

The ability to generate HSCs reliably and efficiently *in vitro* would be incredibly valuable as a therapeutic tool to treat various hematopoietic malignancies and disorders, as a model of various diseases, and as a scientific tool to explore hematopoietic development[54,89]. This ability remains elusive, but, given its immense promise, many researchers are actively developing and refining methods to this end. Two major strategies for *in vitro* hematopoietic differentiation from hPSCs exist. In the first, hPSCs are co-cultured with stromal cells in the presence of serum and cytokines[90]. While this method does generate a variety of hematopoietic cells, it does so with low-efficiency, and is dominated largely by primitive or EMP-derived lineages[91]. The second

method uses a directed differentiation approach, providing the signals that are present during normal development[92]. This method benefits from its ability modulate signaling pathways to derive purely primitive or purely definitive cell populations from a given culture[93,94]. Using this method, the authors were able to generate and identify a hemogenic endothelial population that gave rise to all of the hematopoietic cells in their culture[56]. Unfortunately, no *in vitro* differentiation protocol has been able to generate and maintain HSCs. However, the fact that the directed development approaches are able to undergo definitive hematopoiesis, as evidenced by the presence of T-lymphocytes[93], and generate cells from other hematopoietic lineages, strongly suggests that some of the cells in these cultures pass through an HSC-like stage. The features of the latter system make it the best option for modeling IL through the introduction of specific lesions observed in patients.

2. Early Sequencing

2.1 Introduction

In an initial effort to explore the role of germline variation in the development of IL, we obtained germline DNA samples from a cohort of IL patients and their mothers. Much work had been done focusing on KMT2A-R+ leukemia, and there was speculation that this single somatic event was a sufficiently powerful driver that it might greatly decrease the necessity for other genetic damage[35]. Thus, to maximize the chance of finding germline variants that contribute to the development of IL, we restricted this initial study to KMT2A-R- leukemia.

2.2 Methods

2.2.1 Patient information and DNA samples

DNA samples and demographic and clinical information were collected from 23 pairs of deidentified Caucasian mothers and their infants with KMT2A-R- acute leukemia who were enrolled on the COG-AE24: 'Epidemiology of Infant Leukemia' protocol. Briefly, infants (<12 months) with a confirmed diagnosis of ALL or AML during the period 1996–2006 at North American COG institutions were eligible for the parent AE24 study; cases with Down syndrome were excluded. None of the infants included in this study were reported to have birthmarks, birth defects, known chromosomal abnormalities or family histories of pediatric cancers. In addition to providing buccal cell samples for themselves (via mouthwash) and their infants in first remission (via cytobrushing) using Puregene Buccal Cell Kit (Gentra Systems, Minneapolis, MN USA), as well as consent for genetic research using the samples, mothers also released their child's diagnostic information, including results of Southern blot, reverse transcription-PCR, fluorescent in situ hybridization or other cytogenetics testing, to permit central review. Three independent

reviewers evaluated submitted materials to confirm diagnoses and classify the leukemia as KMT2A-R+, KMT2A-R- or indeterminate. Institutional Review Boards at the University of Minnesota Coordinating Center (#0309M52104) and participating COG institutions approved the parent AE24 study. Control pediatric exomes were obtained from Caucasian infants and their parents without cancer collected as part of an exome sequencing initiative conducted by the Newborn Medicine Division at St Louis Children's Hospital (courtesy of F Sessions Cole, MD). Exome sequencing was approved by Washington University Human Research Protection Office ID# 201105062.

2.2.2 Exome sequencing and data analysis

For all samples, 15–25 ng of germline DNA was whole-genome amplified using the Sigma GenomePlex kit according to the manufacturer's protocol (Sigma, St Louis, MO, USA). From each amplified product, 1 μg was used for sequencing library preparation according to the Illumina TruSeq DNA Sample Prep v2 kit followed by hybridization capture of each exome according to the Illumina TruSeq Exome Enrichment Kit (Illumina, San Diego, CA, USA). Libraries were sequenced three/lane on the Illumina HiSeq 2000 platform generating 101 bp paired-end reads by the Genome Technology Access Center at Washington University.

For all exome data from probands, mothers and controls, we used a published bioinformatic pipeline[95] with sensitivity of 96.9% and specificity of 99.8% with exome analysis for raw data alignment and variant calling. Raw sequence data in fastq format were aligned to the NCBI human genome build 37 (hg19) using a purchased, multi-threading version of Novoalign version 2.05 (www.novocraft.com) and published thresholds. An alignment threshold of 200 was used (-t 200), with adapter stripping (5′-a AGATCGGAAGAGCG-3′) and quality calibration enabled (-

k). Reads with multiple alignments were discarded (-r none, -e 1) and output was in SAM format (-o SAM). Variant calling from the aligned output for the individual exomes was then performed using SAMtools[96]. The aligned data were converted to BAM format to allow the removal of duplicate reads using Picard 'MarkDuplicates'. Variants were then called with the SAMtools version 0.1.18 mpileup command, using options -AB –ugf and bcftools 'view' with settings -bvcg. Finally, variants were filtered with vcfutils 'varFilter' using default settings except retaining all variants with under 99999 reads. This process ultimately yielded a comprehensive list of exomic variants for each subject, including single-nucleotide variants and short insertions and deletions. 'Raw' variant calls from each sample were further filtered by retaining only variants with greater than or equal to5-fold coverage/allele (>10-fold/base position), a genotype quality score of greater than or equal to 10 and a mapping quality score of greater than or equal to60. Although greater than or equal to 5-fold coverage/allele was a bare minimum, it should be noted that our average coverage per variant per exome was 21.5-fold/allele (43-fold/base position). Each of the individual quality score thresholds will only retain a variant position with at least a 95% likelihood of being a true variant. When applied together, the probability of a variant miscall is significantly reduced. All remaining variants were used as input for the 'variants_reduction.pl' tool provided with the ANNOVAR software package (http://www.openbioinformatics.org/annovar/)[97]. To enrich for high-confidence variants likely to confer a functional consequence, successive filters were applied, keeping only variants which were non-synonymous and coding or at splice junctions, and were rare (present at <1% minor allele frequency) in either the 1000 Genomes Project (April 2012 release) or in the dbSNP130 Non-Flagged variants lists. Sequencing results are available at the NCBI Short Read Archive under accession number SRP024273.

2.2.3 Candidate Gene Selection

Using version 63 (ALL) or version 64 (AML) of the COSMIC database (http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/)[98], we compiled lists of genes relevant in AML and ALL. To do this, we used the Tissue search feature, selecting samples from 'hematopoietic and lymphoid' tissue followed by 'NS' for Subtissue type. For AML, we further refined our gene list by selecting 'hematopoietic' from the Histology menu and, from the subHistology menu, 'Acute myeloid leukaemia', 'Acute myeloid leukaemia associated with MDS', 'Acute myeloid leukaemia myelodysplastic syndrome therapy related NOS' and 'Acute myeloid leukaemia therapy related'. For ALL, we selected 'Lymphoid Neoplasm' from the Histology menu followed by 'Acute lymphoblastic leukaemia' and 'Acute lymphoblastic B cell leukaemia' from the subHistology menu. Having filtered by tissue and histology, we selected all genes with sequence variation in our cohort, which yielded a list of 126 ALL-associated genes and 655 AML-associated genes. Thirty-four genes were shared between candidate gene sets. These genes are listed in Table 2.1.

2.2.4 Hypergeometric and permutation testing

Hypergeometric (Fisher's Exact) testing was performed using the 'phyper' function in the R software statistics package (version 2.15.3; available at http://www.r-project.org). P-values (using an α=0.005 to increase stringency) were generated by comparing the aggregate number of rare, non-synonymous, predicted deleterious variation in each patient group against either the matched mothers, the unaffected control population, the opposite patient group or the opposite group of mothers. Unaffected controls consisted of 12 unaffected Caucasian pediatric exomes.

Permutation analysis was executed in the R software package using the 'sample' function in the base package. Using this function, a distribution of the number of rare (not listed in dbSNP 135 or the 1000 Genomes Project), non-synonymous variants was created by performing 100 000 iterations of randomly selecting the number of genes identified via filtering (126 ALL-associated genes or 655 AML-associated genes).

2.2.5 Dideoxy sequencing

Confirmatory dideoxy sequencing was performed at Washington University's Protein and Nucleic Acid Chemistry Laboratory using primers listed in Table 2.2.

2.3 Results

Table 2.3 shows the maternal and infant demographics in the ALL and AML subgroups. The infants with AML presented somewhat earlier than those with ALL (5.3 months vs 8.3 months), but otherwise there were no differences between the subgroups. Maternal age also did not associate with phenotype.

Table 2.4 shows that the average amount of congenital coding variation is higher in affected infants than in mothers or unaffected controls. For infants with ALL, our range of 463–3209 (including insertions and deletions) is consistent with the range of 791–1462 single-nucleotide variants per child reported by Chang[99]. To focus on variants more likely to impart a functional effect associated with acute leukemia, we identified 126 ALL-associated genes and 655 AML-associated genes within the COSMIC database. From these candidate genes, we tabulated the number of congenital variants that were rare and non-synonymous. We found an average of 12 variants per ALL patient in the 126 ALL-associated genes and 163 variants per AML patient in the 655 AML-associated genes, both values exceeded the averages of 6 and 132 observed in

ALL and AML mothers, respectively, as well as the 2 and 28 observed in controls (Table 2.5). Rare, non-synonymous variants in infants with ALL or AML were 2.0 and 1.4 times more likely, respectively, to be found in leukemia-associated genes compared with controls (Table 2.5). There was no correlation between the number of rare, non-synonymous and putatively deleterious variants identified and the size of the gene: R^2=0.21 (ALL) and 0.15 (AML). Given the unexpectedly large numbers of variants identified in candidate genes, infants were tested for an enrichment of variation in candidate genes using a hypergeometric test (Table 2.5). We found that, compared with controls, IL patients and mothers demonstrated a statistically significant enrichment of variation within either set of candidate genes. These results suggest that IL patients are indeed enriched for rare, deleterious variation in leukemia-associated genes.

Several factors might lead to bias in the variant distribution of our samples. For example, differing transcript sizes or systematic sequencing error. To address these potential biases, we performed a randomization test wherein we generated a random set of genes that was the same size as the set of variant genes observed in our samples (7808 in ALL and 8422 in AML), and recorded the number of COSMIC genes in each random set (Figure 2.1). We repeated this procedure 100 000 times and found that the randomly generated sets with the same number of COSMIC genes were not observed for either ALL or AML. This supports the conclusion that the observed enrichment was not due to systematic errors and was specific to our patients. Alternatively, we also generated 100 000 random lists of only 126 or 655 genes and recorded the number of variant genes observed in each iteration. Results (not shown) were qualitatively the same as our initial permutation experiment. Results of maternal random permutation testing are shown in Figure 2.2. Maternal exomes also demonstrate, to a lesser degree than infants, an enrichment of rare, deleterious variation in leukemia-associated genes but none of these mothers

had developed leukemia at the time of study enrollment. We also validated our sequencing variant calls that were unique to an individual with additional dideoxy sequencing (Table 2.6). We did not validate variants observed in matched mothers and infants, as such a result by chance would be exceptionally unlikely.

To prioritize genes that may be most relevant to IL and highlight the combinatorial nature of maternal and non-maternal germline variation, we looked for (a) compound heterozygous genes and (b) genes that were the most commonly variant across all patients. We surveyed all genes for compound heterozygotes, where a gene must show at least two rare, non-synonymous and deleterious variants with at least one seen in the matching mother and at least one variant that was non-maternal. We found that every AML infant was a compound heterozygote for only two genes: KMT2C and ANKRD36. ANKRD36 (ankyrin repeat domain 36) was not a leukemia candidate gene and has not been previously associated to leukemia. We focused on KMT2C because it was in our AML-candidate gene list and owing to its direct connection to leukemia. Interestingly, despite the fact that KMT2C was not on our ALL-candidate gene list, we found that 50% of infants with ALL were also compound heterozygotes. Within KMT2C, we identified nine stop gain variants (Figure 2.3). Six of these were caused by a known, rare T insertion at chr7:151945072 (rs150073007) and three of these were seen in the matching mothers (four of nine total). For other candidate genes, 67% of AML patients were also compound heterozygotes for RYR1 and FLG, whereas 50% of ALL patients were compound heterozygotes for RBMX. We next plotted the top 50 variant candidate genes for infants (Figure 2.4). We found the most variant (but not necessarily compound heterozygotes) AML-associated genes in infants with AML were TTN, KMT2C and FLG (Figure 2.4, columns 1,3,4; Infants: AML), but from the ALL-associated gene list, MDN1, SYNE1 and KMT2B (Figure 2.4, columns 1,2,3; Infants:

AML) were frequently variable. For infants with ALL, we found that MDN1 was the most variable ALL-associated gene (Figure 2.4, column 1; Infants: ALL), but also noted frequent variation in TTN, RBMX and KMT2C from the AML candidate list (Figure 2.4, columns 1,2,3; Infants: ALL). Individual variants and their observed frequencies for the top three most frequently variant ALL and AML-candidate genes are listed in Tables 2.7 and 2.8. Consistent with the hypothesis of a combinatorial inheritance of functionally significant variation in leukemia-associated genes, infants generally show greater variability than mothers. We also observed that infants with AML tend to have more variants across the top genes than ALL infants.

2.4 Discussion

Clearly, a critical component of IL etiology remains undiscovered. The search for these additional insults has been ongoing for decades and has mainly focused on the acquisition of additional somatic mutations within the pre-leukemic clones due either to (a) enhanced mechanisms of mutagenesis from the initial genetic defect or (b) environmental exposures to toxins that promote DNA mutation. Assuming that the typical cancer requires 2–8 mutation in genes regulating cell fate, cell survival and/or genome maintenance[17], neither of these mechanisms appears sufficient to account for the incidence of IL, and genome-sequencing results from KMT2A-R+ infant ALL reported exceptionally few somatic mutations in these three classes of genes[22]. One category of genetic variation that has not been deeply explored in these patients is rare germline variants. The aptly named model of ‘clan genomics’ by Lupski and colleagues[100] posits that combinations of rare and private alleles, in the right genomic context, can combine to exert profound, but variable, influence on complex phenotypes. Considering this

model, we hypothesized that profiles of rare, coding germline variation may comprise some proportion of the expected functional variation typically observed in cancer, but as of yet, not observed in IL under a model of somatic mutation. Under this model, each parent possesses a partial profile of variation, individually insufficient to significantly increase cancer risk, but through random segregation these alleles align in an offspring and result in the right context to significantly increase that child's risk of early leukemogenesis. A recent genome-wide association study found support for an additive model of common variants influencing standard-risk pediatric ALL and proposed that additional low-risk and very rare variants are likely to be present and exerting substantial effects on ALL risk[101]. The Rare Variant Hypothesis posits that a singular complex phenotype may demonstrate a wide variety of functionally significant genetic variants in critical genes or metabolic pathways[102]. Support for this hypothesis is provided by recent studies asserting that genetic variance for complex traits is mostly additive in nature[103]. These congenital profiles of variation may consist of very rare variants of strong effect that may be augmented or modulated by additional low-risk variants, which is consistent with reports describing how multiple functional variants are required for a normal cell to undergo malignant transformation[104].

Although KMT2A-R+ IL has been intensively studied, there are few studies of KMT2A-R− IL and none simultaneously characterizing large-scale maternal sequencing. To our knowledge, this is the largest sequencing survey of KMT2A-R− IL. In exome sequencing of non-cancer DNA from matched mothers and their infants who developed acute leukemia, we find a statistically significant excess of rare, non-synonymous and predicted deleterious sequence variants in genes already known to be mutated in hematologic malignancies in the COSMIC database. In addition, mothers demonstrated enrichment in candidate gene variation over the control population

supporting the interpretation that the observed enrichment in infants is a chance occurrence resulting from the independent segregation of multiple rare variants inherited from each parent, who individually possess a lesser enrichment of variation in the genes in question. Therefore, consistent with existing models of carcinogenesis[17,103], leukemia can only arise after a discrete threshold of deleterious functional variation is surpassed, whether inherited or acquired. Because paternal, sibling and patient leukemia DNA were unavailable in the AE24 study, our ability to distinguish inherited variation versus *de novo* mutation, identify potentially more penetrant combinations of inherited variants and correlate these patterns with the presence of any somatic mutation is limited. However, the patients in our survey were part of an epidemiologic study that failed to identify significant in utero exposures accounting for their IL[105]. A recent review of *de novo* mutation rates in autistic spectrum disorders reports that only one *de novo* mutation per exome is observed in cases that are significantly enriched for de novo mutations[106]. Thus, although we cannot distinguish paternal variation from *de novo* mutations, elevated rates of *de novo* mutation are insufficient to account for the overall enrichment of variation in candidate genes we have identified in this survey. Despite these limitations, our results continue to support our hypothesis that these infants possess germline variability in leukemia-associated genes and pathways that may reduce the amount of functional somatic mutation typically observed in other cancers.

Given the large number of variants observed, particularly in AML patients, we are not suggesting that every variant identified is involved in leukemogenesis, nor that acquired chromosomal rearrangements or somatic mutations are irrelevant. We are providing evidence that these infants harbor an abundance of congenital and putatively functional variation that may drive or modulate early leukemogenesis. Figure 2.4 qualitatively depicts that approximately three to five genes are

commonly variant in the germline of most or all of these infants, consistent with the functional classes of genes and number (two to eight) of variants, thought necessary for carcinogenesis[17]. This model of germline variation influencing leukemogenesis could explain the short latency that has proven difficult to replicate in animal models. The model would not exclude potential leukemogenic effects of topoisomerase II inhibitor exposure, although this has only been associated with KMT2A-R+ IL, and would also be consistent with the lack of heritability observed within pedigrees if co-segregation of many alleles were necessary to predispose to malignancy. A recent Brazilian study of leukemia in children younger than 2 years found a statistically significant increase in adjusted odds ratios of 1.66 for infants with ALL and a near-significant increase of 1.54 for AML when any second-degree relative had cancer, supporting the conclusion of more subtle familial genetic susceptibility in IL[107]. Interestingly, the odds ratios increased significantly when the children's father had any cancer (1.80 ALL and 2.34 AML), but no such significant increase was observed when mothers had cancer.

Additive germline variation could also explain the very high concordance observed in monochorionic twins as both would have the same profile of inherited genetic variation, as well as the relative lack of disease concordance between dichorionic twins despite an 8% incidence of shared placental circulation. The 'intraplacental metastasis hypothesis[108]' is useful to describe the exceptionally high rate of leukemia concordance for monochorionic twins who share intra-placental anastomoses. The blood-borne nature of these hematologic malignancies would also explain why other pediatric cancer types do not show similarly high concordance in monozygotic, monochorionic twins[109]. However, the majority of twins are dichorionic, and approximately 8% demonstrate blood group chimerism because of placental fusion allowing blood exchange[110]. Despite this frequency and multiple reports of twin–twin transfusion

syndrome in dichorionic twins[111–114], we found only one report of concordant IL in dichorionic twins because of the leukemia clone passing between infants, not through inter-placental anastomosis, but through the maternal circulation and not resulting in leukemia in the mother, only the other twin[115]. Although discordant IL in monozygotic twins is rare, Brown and colleagues documented the apparent spontaneous resolution, potentially through an immune-mediated process, of a KMT2A-ENL+ clone from a co-twin of an affected twin[116]. These observations further support the conclusion that these circulating leukemic clones require additional factors in order to proliferate.

KMT2C, a homolog of KMT2A, maps to 7q36, which is a chromosomal region often deleted in myeloid leukemias[117]. Like its KMT2 family members, KMT2C is a H3K4 histone methyltransferase that regulates gene expression through the FYR and SET domains[118]. We identified nine rare or novel germline frameshift insertions that introduces a premature stop codon truncating the C-terminal FYR-N, FYR-C and SET domains (Figure 2.4) necessary for proper target gene expression (for example, HOX), critical for embryogenesis and development[119]. KMT2C has not been previously linked to IL, but has been associated with a variety of solid tumors and did show an enrichment of mutations in adult AML patients[120,121] and germline KMT2C variation was recently reported in a pedigree with adult-onset AML and colorectal cancer[122]. In addition to KMT2C, TTN was frequently variable in these patients' germline and has been previously found to carry somatic mutation and germline variation in multiple cancer types[120,123]. Although best studied during embyonal cardiac development during mesoderm differentiation between heart and blood, TTN has also been shown to be required for proper chromosome packaging and remodeling during cell division[124]. It is reasonable to hypothesize that aberrant chromatin remodeling, either through dysfunctional KMT2C alone or

in concert with dysfunctional TTN during embryonal mesodermal differentiation, may have a crucial role in the etiology of IL in these KMT2A-R− cases.

These data raise interesting new insights into the genetic architecture of KMT2A-R− IL. Future work will focus on additional sequencing of nuclear family pedigrees with an affected infant to further refine the candidate genes influencing leukemogenesis, as well as functional analyses in patient-derived myeloid precursors of profiles of additive variation in the context of KMT2C and TTN dysfunction within iPSC-based in vitro and in vivo model systems.

Figure 2.1: Permuation Testing in IL Exomes. Random permutation testing of gene subgroups in IL ALL and AML patients. The distribution of novel, non-synonymous, deleterious variants in each figure was generated by randomly selecting either 126 (ALL infants) or 655 (AML infants) genes from the patient exomes. The red dot in each panel marks the actual variation observed in each patient group (ALL P=3.6 e-5; AML P=1.0 e-38) from each COSMIC candidate gene set.

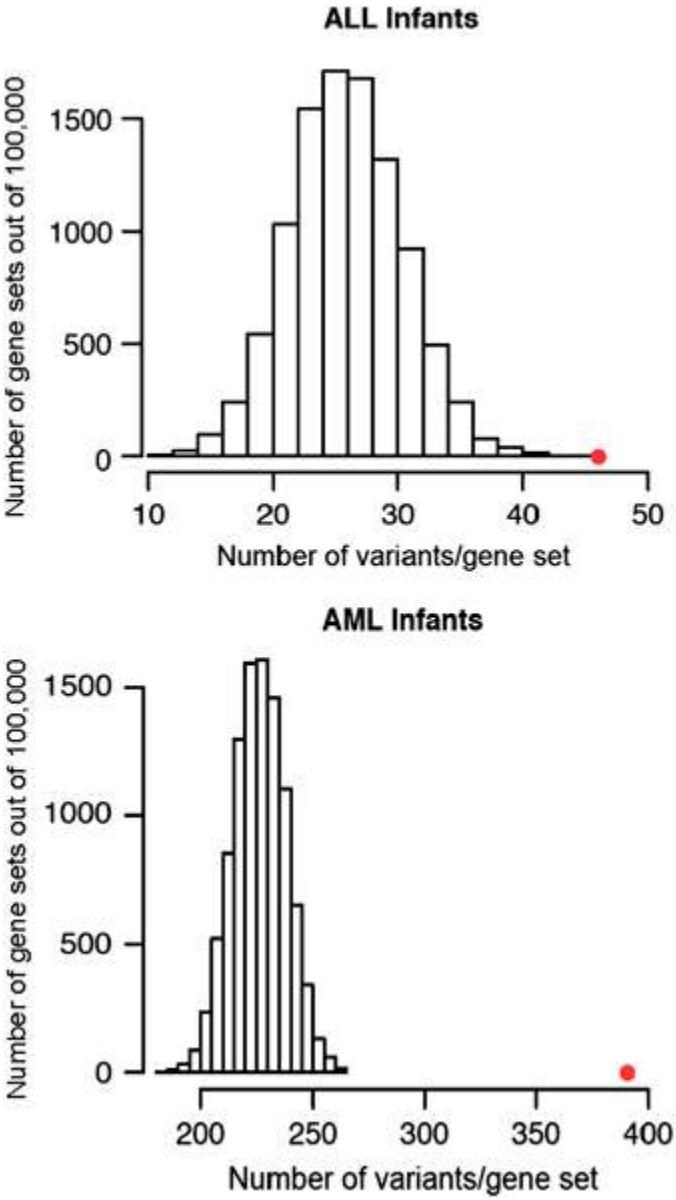


Figure 2.2: **The Top 50 Variant ALL and AML Genes in Infants and Mothers**. Each row indicates an individual and the row position indicates matched pairs (e.g. the top row for ALL infants is the child matched to the mother in the top row of ALL mothers). A colored square indicates a novel, non-synonymous, deleterious variant in that individual in that gene. The shading for each box indicates the number of variants according to the key under the images

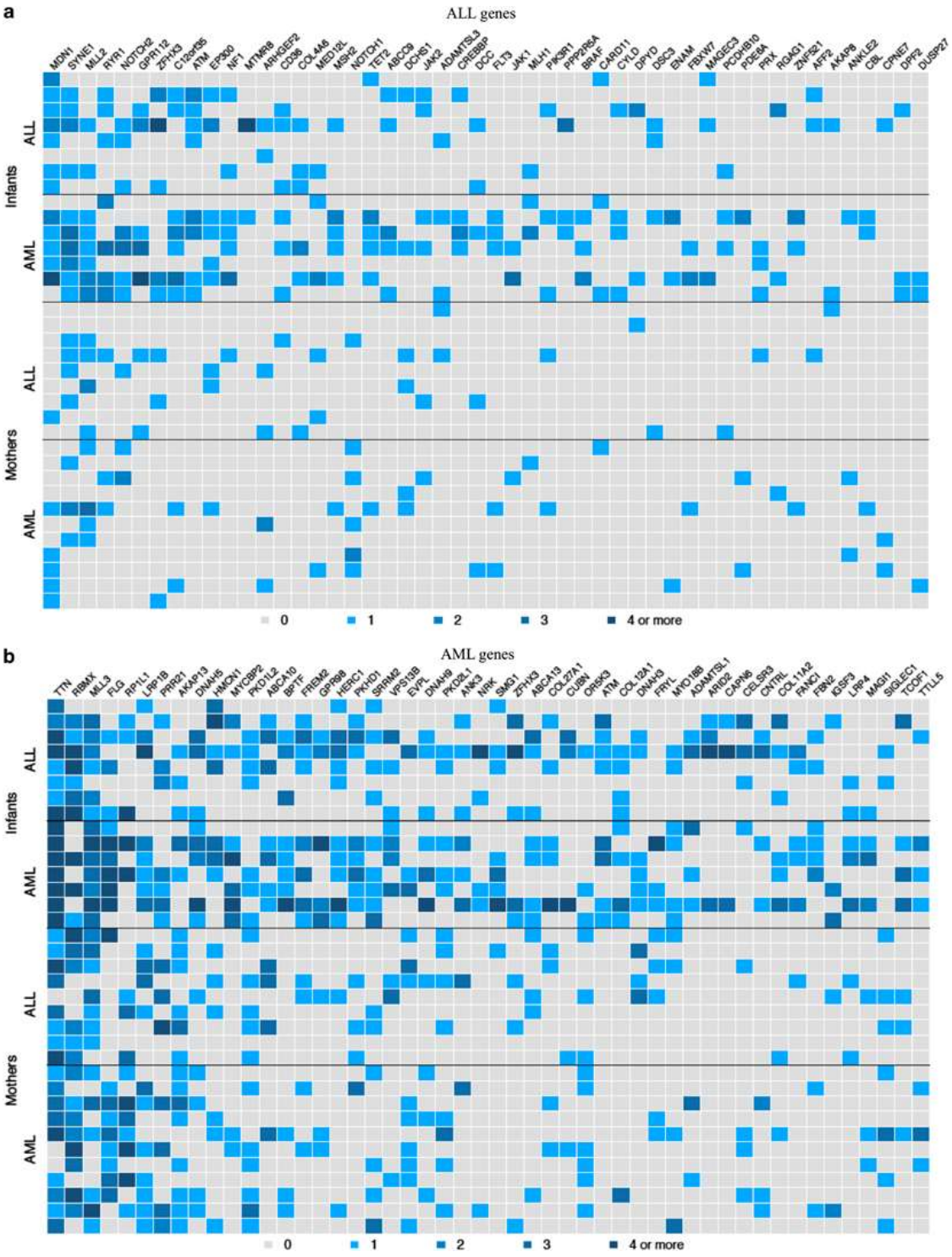


Figure 2.3: Permutation Analysis in Maternal Exomes. Random permutation testing in maternal exomes. The distribution of rare, non-synonymous, deleterious variants in each figure was generated by randomly selecting either 126 (for ALL) or 655 (for AML) sets of genes in 100,000 iterations from the maternal exomes. The red dot in each panel marks the variation observed in each maternal group from each COSMIC candidate gene set. Maternal exomes demonstrate an enrichment of variation in leukemia-associated genes similarly, but to a lesser degree than their infants with acute leukemia.

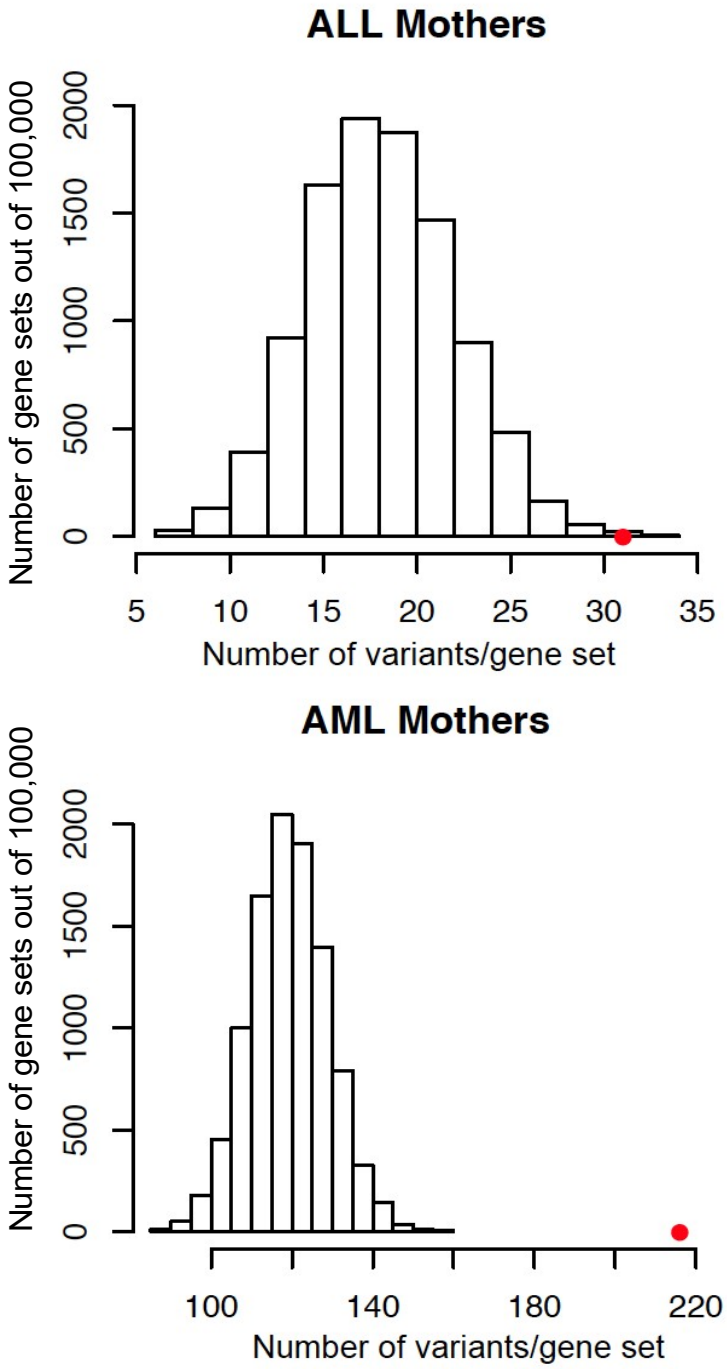


Figure 2.4: Plot of KMT2C Functional Domains and Positions of Germline Sequence Variation. The bar represents the coding sequence of the KMT2C open reading frame with color-coded functional domains in their relative positions. The circles above the bar represent base positions where either missense (blue) or nonsense (red) variants were identified by exome sequencing. Nine infants possessed rare nonsense variants (six at the same base position, rs150073007). Four of these nine nonsense variants were also observed in the matching infants' mothers

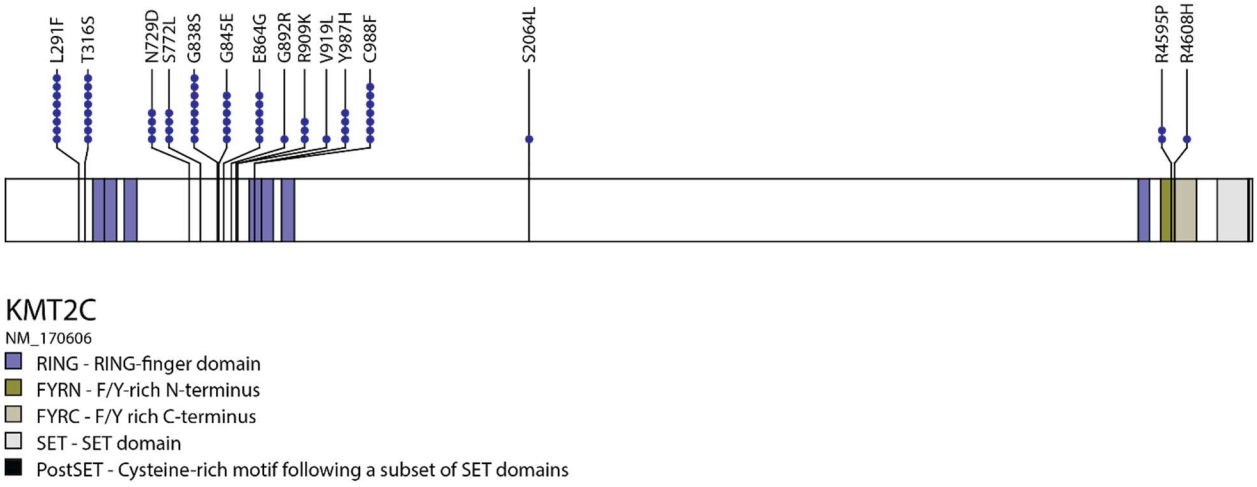


Table 2.1: COSMIC defined candidate genes for AML and ALL.

ALL Candidate Gene List

ABCC9	CARD11	CTNNB1	ETF1	IDH1	MEGF10	OFD1	PRX	STOML2	WDR88
ADAMTSL3	CBL	CYLD	EZH2	IGSF5	MLH1	OR51D1	PTEN	STRADA	ZFHX3
ADCK1	CCND3	DCC	FBXO31	IKZF1	KMT2D	OR8H3	PTPN11	SYNE1	ZNF311
AFF2	CD36	DCHS1	FBXW7	IL7R	MSH2	PAX5	RASGEF1A	TET2	ZNF394
AKAP8	CD74	DENND3	FKBP9	JAK1	MTMR8	PCDHB10	RB1	TLL2	ZNF521
ANKLE2	CD79B	DHX15	FLT3	JAK2	MYD88	PDCD4	RGAG1	TSPAN7	ZWILCH
ARHGEF2	CDC42EP1	DPF2	GBP6	KLF2	MYOM2	PDE6A	RUNX1	TMEM30A	
ATM	CDKN2A	DPYD	PAXBP1	KRAS	NCEH1	PHOX2A	RYR1	TNFAIP3	
B2M	CDKN2C	DSC3	GIMAP5	LRP1B	NEMF	PIK3CA	SCNN1A	TP53	
BCL2L10	COL4A6	DUSP27	ADGRG4	MAGEC3	NF1	PIK3R1	SERPINA1	TRAF3	
BRAF	CPNE7	DUSP9	HLA-DMB	MDN1	NOTCH1	PMS1	SERPINA6	TSC22D1	
BRSK1	CREBBP	ENAM	HMGB1	MED12L	NOTCH2	PPP2R5A	SMARCB1	TSPAN7	
CAMTA1	CRLF2	EP300	HNF1B	MEF2B	NRAS	PRDM1	SOCS1	UBE2A	

AML Candidate Gene List

ABCA10	CACNA1E	DCAF8L1	FBXL18	IDH1	LRRC47	NOS1	PTCH1	SLC17A3	TOP3B
ABCA13	CACNA2D3	DCHS1	FBXL7	IDH2	LRRN2	NOTCH1	PTEN	SLC24A3	TP53
ABCB11	CADM2	DCLK1	FBXO11	IFIH1	LRWD1	NOTCH2NL	PTPN11	SLC25A11	TP53I11
ABCD2	CALB2	DCT	FBXO27	IGHMBP2	LTA4H	NOTCH4	PTPRE	SLC25A12	TP73
ABCG8	CALD1	DCTN1	FBXO3	IGSF21	LUZP2	NOX3	PTPRG	SLC25A20	TPTE2
ABL1	CAMTA1	DDHD1	FFAR1	IGSF3	MAGEB1	NPM1	PTPRN	SLC26A2	TRAM1L1
ASIC1	CAPG	DDX1	FGL2	IKZF1	MAGI1	NRAS	PTPRN2	SLC30A6	TRIM24
ACSS3	CAPN6	DDX60	FILIP1	IL17RD	MAGI2	NRK	PTPRT	SLC37A2	TRPC1
ACTA2	CAPS	DENND2A	FKTN	IL1R1	MAP1B	NRXN2	PTX4	SLC4A11	TRPC4AP
ADAM11	CARD11	DHX34	FLG	IMP4	MAP2	NSD1	PXDNL	SLC51A	TRPM4
ADAM33	CASQ1	DIS3	FLRT2	IMPG2	MAP3K15	NUBP2	RAB17	SMAD4	TSPYL5
ADAMTSL1	CBL	DLGAP3	FLT3	INS	MAPK1	NUMA1	RAB25	SMARCB1	TTC39A
AFF2	CCND3	DLX3	FNDC1	INS-IGF2	MASP1	NYX	RAB36	SMC1A	TTLL10
AGAP1	CDC42	DNAH3	FOXP1	ITGA8	MAX	TENM2	RAD21	SMC3	TTLL2
AGBL1	CDH18	DNAH5	FREM2	ITGAD	MBLAC1	TENM3	RAI2	SMG1	TTLL5
AHCYL1	CDH24	DNAH9	FRMD8	ITGAX	MDFI	OVGP1	RANBP2	SOCS1	TTN
AKAP13	CDH4	DNAJA3	FRMPD3	JAK1	MED14	P2RY2	RASSF7	SOHLH1	TUFT1
ALS2CR11	CDHR2	DNM2	FRYL	JAK2	MEGF11	PA2G4	RB1	SON	U2AF1
AMOT	CDKN2A	DNMT3A	FRZB	JAK3	METTL3	PAMR1	RBKS	SORCS3	U2AF1L4
ANG	CEACAM19	DOCK2	FZD1	JAM2	MGLL	PAPPA2	RBM41	SOS1	UGCG
ANK3	CEBPA	DOCK4	G6PC	KDM8	MIER3	PARD3	RBM46	SOX5	UGT1A10
ANKRD13A	CECR2	DOCK9	GABRG3	KAT2B	MIS12	PARP2	RBMX	SPATA20	UNC5B
ANKRD26	CELSR1	DOK2	GABRR1	KCNA3	MKRN3	PARP6	RCAN2	SPATS1	UQCR10
ANPEP	CELSR3	DRD2	GANC	KCNH5	MKX	PCDHA13	RET	SPTBN5	USP44
AP1G2	CEP128	DROSHA	GATA1	KCNK6	CENPU	PCDHA4	RFC1	SSC4D	USP9X
AP1M1	CEP170	DSCAM	GATA2	KCNQ2	MLIP	PCDHA6	RGS8	SRRM1	VCAM1
APOL6	CHIT1	DYNC1H1	GATAD2B	KCNQ3	KMT2C	PCDHA8	RIMS1	SRRM2	VCAN
ARC	CHRNA4	DYSF	GBP4	KCNT1	MORC3	PCDHB1	RIN1	SRSF2	VIL1
ARHGAP39	CILP2	E2F8	GCAT	KCNT2	MPL	PCDHB10	RIPK4	SRSF6	VIP
ARHGAP5	CLCN6	EED	GDF10	KDM6A	MPND	PCDHB11	RNASE9	SSX7	VIPR1
ARHGEF2	CLVS2	EFL1	GDF5	KHK	MRPL14	PCDHB7	RNF111	STAG2	VPS13B
ARID2	CNGA1	EGFL8	GDPD4	GLTSCR1L	MSH6	PCDHGA2	ROBO2	STC2	WAC
ARSF	CNPPD1	EIF2D	GDPD5	KIAA1217	MSR1	PCDHGC4	RP1	STK32A	WDR11
ASAP2	CNTLN	EIF4B	GLE1	NWD2	MTMR8	PCMTD2	RP1L1	STK36	WEE1
ASTE1	CNTN5	ELANE	GLI1	ARFGEF3	MTUS2	PDGFRA	RPL27A	STRN	WLS
ASTL	CNTNAP4	ELF1	GLRA1	KANSL1	MYBL2	PDK4	RPL9	STT3B	WNK4
ASXL1	CNTRL	ELFN2	GLTPD2	KIDINS220	MYCBP2	PDLIM7	RSRC2	STX3	WT1
ATG16L1	COL11A2	ELL	GNAI2	KIF2C	MYEOV	PDS5B	RUFY1	STXBP2	XIRP1
ATM	COL12A1	ELN	GNB1	KIT	MYH14	PDXDC1	RUNX1	SULT1C2	ZC3H12D
ATP1A2	COL19A1	EML4	GPC2	KLHL29	MYLK2	PGLYRP2	RXFP1	SUPT5H	ZC3H18
ATP1B4	COL27A1	ENC1	GPR183	KLK7	MYO18B	PHF6	SAA4	SUSD5	ZC3H8
ATP2A2	COL2A1	EPB41L5	ADGRV1	KRAS	MYO1F	PHF8	SAP130	SUZ12	ZDHHC11
ATP6V1G3	COL7A1	EPHA8	GPRC6A	KRT1	MYOC	PKD1L2	SCARB1	SV2A	ZFHX2
ATP9B	COL9A1	EPHB1	GRIK2	KRT14	NAE1	PKD1L3	SCEL	SYTL4	ZFHX3
B4GALNT1	COMMD5	ETF1	GRIN2B	KRT79	NALCN	PKD2L1	SCML2	TACR3	ZMYND8

BAAT	CP	ETV6	GRIP1	KSR2	NANOS2	PKHD1	SCUBE3	TBC1D4	ZNF211
BABAM1	CPLX4	EVPL	GRM8	L1CAM	NAPA	PLCH2	SDK2	TBX5	ZNF213
ADGRB1	CREBBP	EWSR1	GSTM3	LAMA5	NAT8L	PLCZ1	SEMA3A	TCEAL3	ZNF236
BBS7	CSF1R	EXOC2	GUCA1A	LAMC1	NAV1	PLIN1	SEMA7A	TCOF1	ZNF260
BCL11B	CTCF	EXOC4	GUCA2A	LARP4B	NCAPH	PLRG1	14-Sep	TCTE3	ZNF276
BCLAF1	CTSG	EZH2	HERC1	LCE1B	NCOA7	PLXNA3	SEZ6L	TET2	ZNF324B
BCOR	CUBN	EZR	HIVEP1	LIMA1	NCR1	PNPLA7	SF3B1	THEG	ZNF34
BECN1	CUL3	AMER2	HJURP	LIPC	NDST3	POLR2A	SFXN2	TRIM24	ZNF37A
BICD1	CYBB	FAM13A	HK1	LIX1	NDUFA13	POU6F2	SH3TC2	TKTL1	ZNF43
BMP5	CYLC1	FAM171A1	HKDC1	LONRF1	NEFH	PPP1R9A	SHANK1	DCSTAMP	ZNF462
BMPER	CYP1A2	FAM19A4	HMCN1	LPL	NEMF	PRDM9	SHQ1	TMEM132C	ZNF485
BMS1	CYP4F8	FAM27E5	HNRNPUL1	LRIG3	NES	PRICKLE3	SHROOM2	TMEM151B	ZNF616
BOC	DAAM2	FAM57B	HOOK3	LRP1B	NF1	PRODH2	SI	TMEM169	ZNF677
BPTF	DAGLA	BRINP3	HRCT1	LRP4	NF2	PRPF40B	SIGLEC1	TMEM198	ZNF687
BRAF	DAGLB	FAM69A	HTR1A	LRRC2	NLRP4	PRPF8	SIMC1	TMTC2	ZNF689
BRPF1	DAOA	FAM83B	HTR3C	LRRC37A3	NLRP8	PRR13	SKOR1	TNKS1BP1	ZNF75D
BRWD3	DAXX	FANCI	HTR5A	LRRC37B	NMNAT2	PRSS27	SLC12A1	TNR	ZNF788
BTBD8	DAZAP2	FBN2	HYDIN	LRRC40	NONO	PSG3	SLC15A1	TNS4	ZRSR2

Table 2.2: Validation of called exome variants via dideoxy sequencing. Fifty rare, single individual variants with ≥5-fold coverage and a quality score ≥20 were selected for secondary dideoxy validation. Of the 50, 43 (86%) were validated and are listed here. Rare variants observed in both mother and child were not chosen for validation as they are unlikely to be false positives.

Gene	Position (hg19)	Subject with Variant	Reference Base	Variant Allele	AA change	Forward primer	Reverse primer
AFF2	chrX:148035259	Infant	C	T	GCA (A) → GTA (V)	cacctgaactgggctgtttt	ctggcaccttcactcacctc
AIF1	chr6:31584079	Infant	G	C	AGG (R) → ACG (T)	ttggtgagaaacgggtgatttgcggg	ttgagcctgtggacaagggtagggat
ALDOA	chr16:30078672	Infant	G	A	GCA (A) → ACA (T)	taccaatatccagcactgaccccgga	tgttctcggtgccaatggactgca
ATXN7L1	chr7:105255128	Infant	C	G	AGG (R) → AGC (S)	atgagggaaagctgtggtgtgggaca	agcaatgccaacccgatgtctcact
COL16A1	chr1:32156158	Infant	G	C	CCA (P) → CGA (R)	cttcacacctggcttcccctgtta	ctcccttcctttgaaacctagaatggct
COPA	chr1:160262988	Infant	T	C	ACT (I) → GTC (V)	ccccttatgggtaggataatttcccttg	cggcagggatgagtggctaatactt
DDX43	chr6:74104671	Infant	G	A	GTT (V) → ATT (I)	tggcacgctactcttacgacgtca	caaaacacagcggcagttcctcgt
EEF1G	chr11:62339051	Infant	C	T	TGG (W) → TGA (stop)	acagacacttcctccatcaactgccc	tttggaggtgagggtgtaggagagga
EFTUD	chr15:82533640	Infant	G	A	GTA (V) → ATA (I)	cccctctttcctcacccttaaccatt	tgggaaaagttggaatggtagctctgg
EXOC3	chr5:465833	Infant	G	C	GGT (G) → CGT (R)	tgaagagggaggttcccagtcaggaa	cgtgcatcccttacctgatgtctgga
FAM196B	chr5:169310673	Infant	G	C	GGG (G) → GCG (A)	actggctgtgagtctcttccttttga	agtgtatggtcctgatccttcccact
GBA2	chr9:35739005	Infant	C	A	GAC (D) → TAC (Y)	ggagccatggttcatcatctgtggga	attccctgcctcccttctgtgtctct
GPR84	chr12:54756704	Infant	G	A	TCA (S) → TTA (L)	tgcaaccattgagccaggtgaggtt	acaggcaagcatccactccaacca
KCNIP4	chr4:20852245	Infant	G	C	CCT (P) → GCT (A)	ctctgggaattgtgtgaaggta	tccgttttcctttgttgctc
KIF11	chr10:94389974	Infant	G	T	CAG (Q) → CAT (H)	actagctagatatcctaccagccagct	tggcagcatcatgaagtttctcctca
LILRA3	chr19:54802548	Infant	G	T	TCC (S) → TAC (Y)	tgtcactgtctgtctctccctccctt	agggaaagtttgtggggaagcctga
NCAPD3	chr11:134062612	Infant	C	T	CTC (L) → TTC (F)	tggaagtcagtcagggaaagagagacca	agatgcgtgcagatccagaaagcct
NCOA2	chr8:71041055	Infant	C	G	CGG (R) → CCG (P)	gtcttagttgatttggctggttctgcac	tccagagccaagcagtagatccaga
OR2V2	chr5:180582262	Infant	GTCT	deletion	Frameshift	ttcttcctcagccagctctccctcat	accttcctcaagccacagtaggggaa
PHKB	chr16:47684830	Infant	C	A	CAG (Q) → AAG (K)	tgaacacagtgagcccttgggaaga	tgctgcaagtgttcaagcatctgact
PIP4K2C	chr12:57992903	Infant	G	A	CGA (R) → CAA (Q)	tccctgggctgtttgtgtatctgct	aaagaacagtcagaacccagcccct
POLR3A	chr10:79773461	Infant	G	T	CTT (L) → ATT (I)	tggcgttgttagttgtggtggtgtt	agtgcccagtattccaggaagccat
PRDM7	chr16:90142285	Infant	G	A	GAA (E) → AAA (K)	ccccattccaatgcacattcagacaga	agtcttcctgctctgaacacccca
PRRC2C	chr1:171509954	Infant	G	C	GTA (V) → CTA (L)	agcaccaattcagccacagtcagtt	ggcctcgacctctgtatccttgatct
PTGES3L	chr17:41123653	Infant	C	T	GAG (E) → AAG (K)	tcctccagagacatgtgtgcagaga	atatgcccacactctcccacacact
RALGAPA2	chr20:20506962	Infant	C	G	CAG (Q) → CAC (H)	cctctgtctggaatactgctggctt	acacagtgaggctagagaagagacagt
RRAD	chr16:66957510	Infant	G	C	CAG (Q) → CAC (H)	ttccctccccgccagttttctttct	tgtgtgttgtgtgttctccaggacgg
SFTPA2	chr10:81317038	Infant	C	T	GGC (G) → AGC (S)	acacactgctcttttccccgacct	tgcagggctccataatgacagtagga
SYNE1	chr6:152469381	Infant	G	A	GAG (L)→GAA (E)	gggatggagtccacactagc	ctagctcccagacgatgagc
SYNE1	chr6:152570334	Infant	C	A	TGC (A)→TGA (stop)	caaacaaaagtgccactgtga	tgagttttcccgttgctttc
SYNE1	chr6:152576794	Infant	C	T	TTC (E)→TTT (F)	aaaagtgctgtggcaacaaa	ggccccatcctgatattttt
SYNE1	chr6:152631566	Mother	G	A	ACG (R)→ACA (T)	gcttacctgccgatgagaga	tgctcacctgtgatgtggtt
SYNE1	chr6:152631653	Infant	C	A	ATC (D)→ATA (I)	gcttacctgccgatgagaga	tgctcacctgtgatgtggtt
SYNE1	chr6:152642398	Mother	G	A	CGC (A)→CAC (H)	tcacacaagaggactgacccta	ccacaaatcaccgacagaac
SYNE1	chr6:152650962	Infant	G	C	TGT (T)→TCT (S)	tcagcgaggctgtgttctaa	acctcaacctgcaggacatc
SYNE1	chr6:152674464	Mother	C	A	CTT (K)→ATT (I)	tgcctacctccaacgtcttc	ctggcacaggcctttacttc
TRPV4	chr12:110230503	Infant	G	C	ACC (T) → AGC (S)	atgtggtgtgtgtgtgactccctcca	tgtcttcccctccagagcctcattgt
TXNIP	chr1:145440753	Infant	G	A	AGT (S) → AAT (N)	tgttaccacagctgtcttgtttctccag	ccacataagaatcctgcccaagaaatgc
UGT3A2	chr5:36039780	Infant	C	G	GAC (D) → CAC (H)	tgggccaatgagaacactgacactt	aggacatgaccacagtgcagctagt
ZHFX3	chr16:72830465	Infant	G	A	TGG (P)→TAG (stop)	ggtgcaattgtaggtgaggtg	tcaagagcacgttcatcagaa
ZHFX3	chr16:72832557	Mother	C	T	TCC (G)→TCT (S)	ctgaccggttgctgattctt	ggtgccatccttgaaacaaa
ZNF362	chr1:33764618	Infant	G	C	GGC (G) → CGC (R)	tccctgttctgaatctcatccctgc	tctggaagattgggaaggtcgtgag
ZNF717	chr3:75788130	Infant	C	T	GGG (G) → GAG (E)	ctctcctgtgtgtgtctgctgatgt	accaacagcaacacatcaactcagga

Table 2.3: Demographic characteristics of the study cohort

		ALL	AML
Sex	Boys	4	6
	Girls	8	7
Avg age at diagnosis (months)		8.3 (0.6 - 11.4)	5.3 (1.6 - 11.4)
Avg maternal age (years)		31.9 (21.3-40.6)	33.4 (25.4-41.8)
No. mothers >35 yrs		3	5

Table 2.4: The average and range of filtered variants per exome in each subgroup. Total exomic variants (single nucleotide variants and INDELs) were filtered for variants that were novel (not previously included in dbSNP 135 and the 1,000 Genomes Project), non-synonymous, with coverage ≥5-fold, a genotype quality score ≥10 and a mapping quality score of ≥60.

	Average total variants per exome	**Range**
ALL infants	1,264.4	463 – 3,209
ALL mothers	1,112.6	985 – 1,267
AML infants	2,549.9	519 – 5,545
AML mothers	1,225.0	1,000 – 1,660
Unaffected controls	582.8	467 – 719

Table 2.5: Hypergeometric analysis of variation in leukemia-associated genes determined by comparing the observed amount of RNS variation in the 126 (ALL) or 655 (AML) COSMIC-identified candidate genes against the expected amount of similar sequence variation observed by randomly selecting 126 or 655 genes from the same patients. P-values generated from hypergeometric (Fisher's Exact) test with $\alpha = 0.005$ (* = not significant).

	Group	**Average variants / exome**	**Range**	***P*-value***
ALL genes (n = 126)	ALL infants	12.1	3 – 33	$3.6\ e^{-5}$
	ALL mothers	6.4	3 – 11	$1.4\ e^{-3}$
	Unaffected controls	1.9	0 – 4	0.24*
	AML infants	22.7	4 – 37	$3.0\ e^{-9}$
	AML mothers	8.2	4 – 16	$1.7\ e^{-9}$
AML genes (n = 655)	AML infants	163.4	38 – 358	$1.0\ e^{-38}$
	AML mothers	132.5	60 – 667	$5.3\ e^{-19}$
	Unaffected controls	27.5	12 – 37	0.007*
	ALL infants	59.4	24 – 131	$5.2\ e^{-29}$
	ALL mothers	49.6	40 – 67	$1.5\ e^{-11}$

Table 2.6: The likelihood of possessing as RNS variant in a leukemia-associated gene. Likelihood ratios were calculated as follows:

$$\frac{\text{Number of rare, non-synonymous, deleterious variants (either ALL or AML mothers or infants) / total number of variants in the respective group}}{\text{Number of rare, non-synonymous, deleterious variants in controls / total number of variants in controls}}$$

Leukemia subtype	Subgroup	Likelihood ratio of having a rare, non-synonymous, deleterious variant in a COSMIC-defined candidate gene relative to controls.
ALL	Infants	2.01
	Mothers	1.62
AML	Infants	1.44
	Mothers	1.44

Table 2.7: Individual variant listings for each RNS variant called variants with ≥5-fold coverage, a genotype quality score of ≥10, and a mapping quality score of ≥60 (see Methods) in the top three most commonly variable genes from the AML-candidate gene list (see Figure 2, Panel B). Each row lists an individual variant by gene, position (hg19), frequency in the infant exomes, frequency in the maternal exomes, and the dbSNP identification number (if applicable, blank means the variant is novel).

Variant list for the top three variant AML-candidate genes (see Figure2, Panel B) - page 1 of 2									
AML Infants					ALL Infants				
Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID	Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID
TTN	chr2:179393490	14%	0%		TTN	chr2:179392343	13%	0%	
TTN	chr2:179395413	14%	0%		TTN	chr2:179393346	0%	10%	
TTN	chr2:179413687	14%	0%		TTN	chr2:179404199	13%	0%	
TTN	chr2:179414436	14%	0%		TTN	chr2:179413565	13%	0%	
TTN	chr2:179414817	14%	0%		TTN	chr2:179422215	13%	0%	
TTN	chr2:179415929	14%	0%		TTN	chr2:179425264	0%	10%	
TTN	chr2:179416935	0%	10%		TTN	chr2:179427497	13%	0%	
TTN	chr2:179417152	0%	10%		TTN	chr2:179429150	13%	0%	
TTN	chr2:179417452	14%	0%		TTN	chr2:179432641	13%	0%	
TTN	chr2:179419226	0%	10%		TTN	chr2:179434267	13%	0%	
TTN	chr2:179421596	14%	0%		TTN	chr2:179436754	13%	0%	
TTN	chr2:179421791	0%	10%	rs183013408	TTN	chr2:179441457	13%	0%	
TTN	chr2:179422214	14%	0%		TTN	chr2:179455560	13%	0%	
TTN	chr2:179424856	14%	0%		TTN	chr2:179458072	13%	0%	
TTN	chr2:179432185	14%	10%		TTN	chr2:179460461	13%	0%	
TTN	chr2:179433832	14%	0%		TTN	chr2:179462635	0%	10%	
TTN	chr2:179434798	14%	0%		TTN	chr2:179474277	0%	10%	
TTN	chr2:179435903	0%	10%		TTN	chr2:179481277	13%	0%	
TTN	chr2:179437494	14%	0%		TTN	chr2:179482201	13%	0%	
TTN	chr2:179439611	14%	0%		TTN	chr2:179485631	0%	10%	
TTN	chr2:179440635	14%	0%		TTN	chr2:179494968	0%	10%	rs192766485
TTN	chr2:179440995	14%	0%		TTN	chr2:179500768	0%	10%	
TTN	chr2:179444512	14%	0%		TTN	chr2:179553427	13%	0%	
TTN	chr2:179444789	14%	0%		TTN	chr2:179566946	0%	10%	
TTN	chr2:179447898	0%	10%		TTN	chr2:179570047	0%	10%	
TTN	chr2:179449116	0%	10%		TTN	chr2:179572323	0%	10%	
TTN	chr2:179453343	14%	10%	rs191549948	TTN	chr2:179582796	13%	0%	
TTN	chr2:179455677	0%	10%		TTN	chr2:179594134	13%	0%	
TTN	chr2:179456083	14%	0%		TTN	chr2:179596188	13%	0%	
TTN	chr2:179458118	14%	0%		TTN	chr2:179599069	13%	0%	
TTN	chr2:179458151	14%	0%		TTN	chr2:179599521	13%	0%	
TTN	chr2:179464371	14%	0%		TTN	chr2:179600264	13%	0%	
TTN	chr2:179466511	14%	0%		TTN	chr2:179604611	13%	0%	
TTN	chr2:179469558	14%	0%		TTN	chr2:179605815	0%	10%	rs201888760
TTN	chr2:179473055	14%	0%		TTN	chr2:179614489	13%	0%	
TTN	chr2:179473455	0%	10%		TTN	chr2:179614883	13%	0%	
TTN	chr2:179476243	0%	10%		TTN	chr2:179615326	13%	0%	rs142848087
TTN	chr2:179476610	14%	0%		TTN	chr2:179632544	13%	0%	
TTN	chr2:179483493	14%	0%		TTN	chr2:179634421	25%	40%	rs200875815
TTN	chr2:179483524	14%	0%		TTN	chr2:179636009	0%	10%	
TTN	chr2:179486004	14%	0%		TTN	chr2:179650587	0%	10%	rs199507913
TTN	chr2:179501253	14%	0%		TTN	chr2:179650627	13%	0%	
TTN	chr2:179504807	14%	0%		TTN	chr2:179664619	13%	0%	
TTN	chr2:179516237	14%	0%		RBMX	chrX:135956408	75%	70%	rs76876438
TTN	chr2:179535890	14%	0%		RBMX	chrX:135956462	50%	50%	rs74463481
TTN	chr2:179549399	14%	0%		RBMX	chrX:135956506	50%	40%	rs77794331
TTN	chr2:179571461	14%	0%		RBMX	chrX:135956573	50%	10%	
TTN	chr2:179575569	14%	0%		RBMX	chrX:135957672	0%	10%	rs139356075
TTN	chr2:179583286	14%	0%		RBMX	chrX:135957690	0%	10%	rs150541875
TTN	chr2:179587640	14%	0%		RBMX	chrX:135957700	0%	10%	rs139954333
TTN	chr2:179590307	14%	0%		RBMX	chrX:135957716	0%	10%	rs142284545
TTN	chr2:179594158	14%	0%		RBMX	chrX:135958704	0%	10%	rs112089728
TTN	chr2:179594648	14%	0%		RBMX	chrX:135958730	0%	10%	rs78702689
TTN	chr2:179594867	14%	0%		RBMX	chrX:135960119	13%	10%	rs76812369

Variant list for the top three variant AML-candidate genes (see Figure2, Panel B) - page 2 of 2

AML Infants					ALL Infants				
Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID	Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID
TTN	chr2:179594933	14%	0%		RBMX	chrX:135960147	50%	20%	
TTN	chr2:179595693	0%	10%		RBMX	chrX:135961560	13%	10%	rs80321628
TTN	chr2:179599124	14%	0%		RBMX	chrX:135961585	0%	10%	
TTN	chr2:179600360	14%	0%		MLL3	chr7:151842305	13%	0%	
TTN	chr2:179605985	14%	0%		MLL3	chr7:151860230	0%	10%	rs142835638
TTN	chr2:179606240	14%	0%		MLL3	chr7:151873435	13%	0%	
TTN	chr2:179610430	14%	0%		MLL3	chr7:151902197	0%	10%	rs138119145
TTN	chr2:179610827	14%	0%		MLL3	chr7:151927016	0%	10%	rs141049734
TTN	chr2:179610922	14%	0%		MLL3	chr7:151927025	13%	0%	rs183684706
TTN	chr2:179610943	14%	0%		MLL3	chr7:151932945	25%	10%	rs199504848
TTN	chr2:179610988	14%	0%		MLL3	chr7:151945072	75%	40%	rs150073007
TTN	chr2:179610989	14%	0%		MLL3	chr7:151945228	0%	10%	rs200184971
TTN	chr2:179611336	0%	10%		MLL3	chr7:151960181	13%	0%	
TTN	chr2:179612343	14%	0%		MLL3	chr7:151962265	0%	30%	rs201834857
TTN	chr2:179613163	14%	0%		MLL3	chr7:151970877	13%	0%	rs138627563
TTN	chr2:179634421	14%	50%	rs200875815					
TTN	chr2:179640696	14%	0%						
TTN	chr2:179642589	0%	10%						
TTN	chr2:179659744	0%	10%						
TTN	chr2:179669360	14%	0%						
RBMX	chrX:135956408	43%	60%	rs76876438					
RBMX	chrX:135956462	29%	60%	rs74463481					
RBMX	chrX:135956506	29%	40%	rs77794331					
RBMX	chrX:135956573	29%	30%						
RBMX	chrX:135960147	14%	0%						
RBMX	chrX:135961560	14%	0%	rs80321628					
RBMX	chrX:135961585	0%	10%						
MLL3	chr7:151841869	14%	0%						
MLL3	chr7:151875022	14%	0%						
MLL3	chr7:151884389	14%	0%						
MLL3	chr7:151891103	14%	0%						
MLL3	chr7:151902304	14%	0%						
MLL3	chr7:151919134	14%	0%						
MLL3	chr7:151927025	43%	20%	rs183684706					
MLL3	chr7:151932945	29%	0%	rs199504848					
MLL3	chr7:151945072	86%	30%	rs150073007					
MLL3	chr7:151945225	0%	10%	rs202098135					
MLL3	chr7:151949795	14%	0%						
MLL3	chr7:151962134	0%	10%	rs146238849					
MLL3	chr7:151962168	0%	10%	rs138908625					
MLL3	chr7:151962265	0%	10%	rs201834857					
MLL3	chr7:151970859	14%	0%	rs149992209					
MLL3	chr7:152012416	14%	0%						
MLL3	chr7:152027753	14%	0%						
MLL3	chr7:152055740	14%	0%						

Table 2.8: Individual variant listings for each rare, non-synonymous, predicted deleterious variant called with ≥5-fold coverage, a genotype quality score of ≥10, and a mapping quality score of ≥60 (see Methods) in the top three most commonly variable genes from the ALL-candidate gene list (see Figure 2, Panel A). Each row lists an individual variant by gene, position (hg19), frequency in the infant exomes, frequency in the maternal exomes, and the dbSNP identification number (if applicable, blank means the variant is novel).

Variant list for the top three variant ALL-candidate genes (see Figure2, Panel A) - page 1 of 1									
AML Infants					ALL Infants				
Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID	Gene	Position	Infant Frequency	Mother Frequency	dbSNP ID
MDN1	chr6:90377741	14%	0%		MDN1	chr6:90363929	0%	10%	
MDN1	chr6:90383180	14%	0%		MDN1	chr6:90363955	13%	0%	
MDN1	chr6:90383935	14%	0%		MDN1	chr6:90368355	13%	0%	rs115792683
MDN1	chr6:90388424	14%	0%		MDN1	chr6:90368471	13%	0%	
MDN1	chr6:90398446	14%	0%		MDN1	chr6:90368490	13%	0%	
MDN1	chr6:90405449	0%	10%		MDN1	chr6:90372565	13%	0%	
MDN1	chr6:90405586	0%	10%		MDN1	chr6:90385883	13%	0%	
MDN1	chr6:90420493	14%	0%		MDN1	chr6:90385922	13%	0%	
MDN1	chr6:90424435	0%	10%	rs150248107	MDN1	chr6:90396621	13%	0%	
MDN1	chr6:90434947	14%	0%		MDN1	chr6:90434940	13%	0%	
MDN1	chr6:90450026	0%	10%	rs114779526	MDN1	chr6:90448168	13%	0%	
MDN1	chr6:90459345	14%	0%		MDN1	chr6:90504494	13%	0%	
MDN1	chr6:90513126	14%	0%		SYNE1	chr6:152461288	13%	0%	
MDN1	chr6:90513189	0%	10%	rs143308656	SYNE1	chr6:152462353	13%	0%	
SYNE1	chr6:152443578	14%	0%		SYNE1	chr6:152462387	13%	0%	
SYNE1	chr6:152563445	14%	0%		SYNE1	chr6:152532645	0%	10%	
SYNE1	chr6:152570334	14%	0%	rs149272010	SYNE1	chr6:152542011	0%	10%	
SYNE1	chr6:152576099	0%	10%		SYNE1	chr6:152576794	13%	0%	
SYNE1	chr6:152577804	14%	0%		SYNE1	chr6:152642378	13%	0%	
SYNE1	chr6:152631566	14%	10%	rs145899734	SYNE1	chr6:152644741	0%	10%	
SYNE1	chr6:152642398	0%	10%	rs140850000	SYNE1	chr6:152658104	13%	0%	
SYNE1	chr6:152688393	14%	0%		SYNE1	chr6:152804248	0%	10%	
SYNE1	chr6:152720877	14%	0%	rs112061681	MLL2	chr12:49425964	0%	10%	rs200315963
SYNE1	chr6:152737745	14%	0%		MLL2	chr12:49426773	13%	0%	
SYNE1	chr6:152763314	14%	0%		MLL2	chr12:49428694	0%	10%	rs146044282
SYNE1	chr6:152771967	0%	10%	rs141464488	MLL2	chr12:49432416	0%	10%	
SYNE1	chr6:152777056	14%	0%		MLL2	chr12:49433599	0%	10%	rs147706410
SYNE1	chr6:152787119	14%	0%		MLL2	chr12:49434934	0%	10%	
MLL2	chr12:49418376	14%	0%		MLL2	chr12:49435457	0%	10%	
MLL2	chr12:49420207	0%	10%		MLL2	chr12:49441813	13%	0%	
MLL2	chr12:49421675	14%	0%						
MLL2	chr12:49425287	0%	10%						
MLL2	chr12:49426100	0%	10%						
MLL2	chr12:49436413	14%	0%						
MLL2	chr12:49443500	14%	0%						
MLL2	chr12:49444031	0%	10%						
MLL2	chr12:49444700	14%	0%						
MLL2	chr12:49445392	0%	10%	rs202076833					
MLL2	chr12:49445967	14%	0%						
MLL2	chr12:49447294	14%	0%						
MLL2	chr12:49448150	14%	0%						

3. Later Sequencing

3.1 Introduction

Our early sequencing results showed a clear enrichment of putatively damaging variation in known leukemia associated genes. The similar, but less pronounced enrichment that was present in the mothers of these infants suggests that there is a threshold of variation that needs to be reached, or that additional insults, likely during a narrow developmental window, are required to turn the observed enrichment of variation into frank leukemia. While the initial sequencing studies that we performed were done as rigorously as possible at the time, there were a number of weaknesses present. The small sample size, an inevitable consequence of the rarity of the disease and our focus on KMT2A-R- leukemia, limited the power of our analyses and the scope of investigation that we could perform. The 1000 genomes database was used for filtering and quality control. At the time, this resource was the best available, but in the years that followed, a number of resources replaced it, most notably the exome aggregation consortium (ExAC) database[125]. Similarly, the best practices for exome sequence data processing, controls and analysis have also matured[126].

In the sequencing results in this chapter, we address these shortcomings, and present a greatly expanded analysis of the germline variation present in IL patients. Importantly, we greatly increase our samples size, sequence controls in the same workflow, employ the powerful ExAC database to filter and curate our variants, and expand beyond candidate gene lists to discover new genes that are more highly variant in IL patients

3.2 Methods

3.2.1 Samples

DNA samples were collected from infants with KMT2A-R+ acute leukemia who were enrolled on the COG-AE24: 'Epidemiology of Infant Leukemia' protocol as described previously in section 2.2.

DNA from an additional 8 IL patients was obtained from the Children's Hospital of Westmead. Sequencing and analysis of these samples was performed as described for the AE24 cohort.

Additionally, 22 full genome sequences from the Pediatric Cancer Genome Project (PCGP), a collaboration between Washington University in St Louis and St. Judes Hospital, were obtained from dbGAP.

As a control, fourteen individual exomes were generated from DNA obtained from the Nurse's Health Study. These were all from healthy adults with no known history of cancer.

Individual level data from 77 participants in the 1000 Genomes Project[127] were used as additional controls in some analyses.

3.2.2 Exome Sequencing

As these sequencing studies were performed over several years, a variety of exome library kits were used. All of these kits were used according to manufacturer instructions. The various kits used are: Illumina TruSeq Nano DNA library preparation kit, Illumina TruSeq DNA Sample Prep v2 kit. Library preparation was followed by hybridization capture of each exome according to the Illumina TruSeq Exome Enrichment Kit (Illumina, San Diego, CA USA) or the IDT xGen

exome research panel. Sequencing was performed on Illumina 2500, HiSeq or NovaSeq machines.

3.2.3 Analysis pipeline

Raw reads were aligned to the hg19 reference genome using bowtie-2 using options –N 1 –X 2000 –p 8. The resulting .sam file is sorted and indexed using samtools[96]. Duplicates are marked and removed using samtools. The GATK indelRealigner is used to realign around spurious indel sites[128]. The resulting .bam file is used as input to samtools mpileup for variant calling using options –q 5 –Q 15. Variants were filtered using awk to retain only the variants called in positions with 5 or more reads, and quality scores >= 10. Further filtering and annotation was performed using the variantsReduction.pl tool from the ANNOVAR software package[97]. This step kept only variants that introduced non-synonymous coding changes, splice site changes, or changes in the UTR. Further, variants present at >0.01 MAF in the ExAC[125] database were discarded. The resulting rare, non-synonymous (RNS) variants were enriched for putatively functional variation. Further manual curation of candidate and target genes was performed to ensure that only the highest quality variants were retained. Specifically, any variant that did not pass the VQSRTranche, Adj_AC, or Inbreeding_Coeff_filter in ExAC were discarded.

3.2.4 Statistical Analyses

An initial analysis that used the same hypergeometric testing for enrichment of deleterious variation in COSMIC genes was performed as previously described.

To observe the differences in the frequency of rare variation in all genes across the exome, the number of variant alleles for each gene in each sample (i.e. IL patients, NHS controls, EVS

samples or ExAC samples) was counted. This number was then divided by the number of alleles sequenced in that sample. The resulting frequencies were plotted using R (version 3.2.1; available at http://www.r-project.org). A linear model of the resulting scatterplot was fit using the lm package in R. All genes that were more than five standard deviations from the regression line were deemed "highly variable." This list of highly variable genes contained a large number of seemingly spurious genes (eg. Olfactory receptors, mucin genes). To ensure that only the genes that were most likely to contribute to IL were retained, this list was filtered to keep only the genes that were "Loss-of-function intolerant" as defined by ExAC, as well as the genes that were not also highly variant in control samples. The resulting gene list was used for downstream analysis.

3.3 Results

This new cohort of IL exomes was much larger than the original focused KMT2A-R- study we performed. As a first pass, and to validate our previous findings, we performed the same hypergeometric test to determine whether there was an enrichment of rare, non-synonymous variation in known leukemia-associated genes. As expected, we observed a marked enrichment regardless of KMT2A-R status or leukemia type (Table 3.1). This was an encouraging finding, especially since our improved filtering strategy had removed many more variants than had been previously. Notably, some of the most frequently observed variants in KMT2C when filtered by 1000 genomes, were found to be spurious when the much larger ExAC database was used as a filter. Despite this, though, there was still a marked enrichment of RNS variants in leukemia-associated genes in all IL subsets, but not in controls.

Given the much greater power afforded by this larger cohort, we next expanded our interrogation from a short list of genes selected *a priori* to the entire genome. To do this we tallied the number of RNS variant alleles in each gene for a given cohort, and divided it by the number of alleles sequenced in that cohort. To ensure that this method was reliable, we first compared this measure of variation for each gene in ExAC to the same measure calculated from the Exome Variant Server (EVS) (Figure 3.1)[129]. As expected, there is a very strong linear relationship between these databases. This analysis makes it clear that while there is a wide range in this measure between genes, it is largely consistent for a given gene between similar populations. The variability between genes will be a function of evolutionary constraint (or lack thereof), gene size, variable mutation rates based on genomic location[130,131] and possibly other factors. These factors, though, will be acting in roughly the same way in different groups, provided these groups are similar in makeup. Hence, when comparing the variation in every allele from the EVS to every allele in ExAC, we see that the variation present in a given gene is relatively constant.

This relationship is largely preserved when comparing between our control exomes and ExAC (Figure 3.2). Due to the much smaller sample size, a few genes appear to have much higher variation in our control exomes than in ExAC. While we expected this to occur due to random fluctuations present in the small sample, we wanted to ensure that this "enrichment" was indeed random. To do this, we applied a filter based on the "Loss-of-function (LoF) intolerance" measure developed by ExAC. Briefly, this scores each gene based on the ratio of the number of possible LoF variants to the number of observed LoF variants. This ratio is calculated for each gene, and genes in the top 10% (i.e. those with the far fewer observed LoF variants than expected) are deemed LoF intolerant. When we filtered the "enriched" genes in the control exomes, we saw that very few of them were LoF intolerant. Further supporting the fact that they

are artifacts, we see that many of them are extremely large genes and enriched for olfactory receptors and mucin genes, all of which tend to have increased artifacts[131].

We then compared our IL exomes to ExAC using the same measure. Again, we see that RNS variation in several genes that are markedly elevated in IL relative to ExAC (Figure 3.3). However, unlike the control exomes, many of these genes are LoF intolerant. Indeed, many of the genes that did not pass the LoF-intolerance filter are shared between the control and IL exomes, further supporting their artifact status. In addition to these artifacts, however, a large number of LoF intolerant genes have far more RNS variation in IL patients than in the ExAC cohort. Presumably, this group is enriched for genes that contribute to the development of IL.

When we explored the genes that comprised this group, a number of patterns emerged. First, KMT2C and KMT2B were both present. This corroborates our finding in the initial KMT2A-R- cohort, and, when considered along with the frequency of KMT2A-rearrangements in IL, makes a strong argument for the critical role of KMT2 genes in infant leukemogenesis. We further explored this idea with a direct interrogation of each KMT2 gene. The KMT2 genes each anchor a complex of proteins known as COMPASS complexes (COMplex of Proteins ASsociated with SET)[132]. Since the function of these gene products will rely, at least in part, on the function of the complexes that they anchor, we also included known members of COMPASS complexes in this analysis. Indeed, there is a significant enrichment of RNS variation in these genes in IL patients regardless of KMT2A-rearrangement status relative to control exomes, 1000 genomes participants and ExAC (Figs 3.4 and 3.5). Interestingly, patients with KMT2A-R+ leukemia had relatively less RNS variation in KMT2A and KMT2B than did patients with KMT2A-R- leukemia. Conversely, KMT2A-R+ patients tended to have less RNS variation in KMT2C and KMT2D than their KMT2A-R- counterparts. While not conclusive, these data are consistent with

the hypothesis that dysfunction of both KMT2A/KMT2B complex genes and KMT2C/KMT2D complex genes is an important feature of IL. This dysfunction can arise through germline variation in either complex, but the requirement for germline variation in the KMT2A/KMT2B complex is abrogated in the presence of a somatically acquired KMT2A rearrangement. Regardless of whether concurrent dysfunction in both of these complexes is a requirement for IL, it seems clear that the KMT2 family of genes play an important role in this disease.

In addition to KMT2B and KMT2C, there were many other genes present in the RNS variant-enriched group (Table 3.2). Interestingly, many of these genes have similar biological functions. Genes encoding ubiquitin ligases, transcriptional modifiers and RNA binding proteins in particular seemed to capture the majority of this group. There were also a number of genes with a previously known link to leukemia (e.g. PTEN). As a group, these genes provide strong clues about the processes that may lead to IL and provide an excellent set of candidates for functional studies.

It is unlikely that any single gene is responsible for the development of IL. Thus, we next turned our attention to the combinations of variants present in each individual. We focused on those genes that are enriched for RNS variation relative to ExAC. As seen in Figure 3.5, all IL patients have at least one variant in at least one of these genes. So while we have potentially captured many of the genes that contribute to infant leukemogenesis, the list presented is likely incomplete. It seems much more likely that combinations of multiple RNS variants in multiple pathways will be required for the development of IL. Interestingly, there seem to be genes that are rarely variant in the same individual. Our sample size is too small to make a strong claim on this point, but the observation is consistent with mutual exclusivity of variation. In this case, certain genes will sufficiently disrupt a pathway or process eliminating the need for any other

variants in this pathway or process. If this holds true, an exploration of the mutually exclusive pathways and processes will yield great new insights into the events that lead to leukemic transformation in infants.

3.4 Discussion

The data presented in this chapter corroborate and greatly extend the findings made in our KMT2A-R- cohort. They confirm that an enrichment of putatively deleterious variation in leukemia-associated genes is a common feature of IL patients. This enrichment has not been observed in any of the control populations that we have sequenced, or in other publicly available control cohorts. Importantly, this observation holds when we use the much more stringent ExAC-based filtering strategy that we have more recently employed. The previous data had a number of false-positives in some of our focus genes. These even persisted when we performed Sanger sequencing on individual variants. The aggregation of thousands of individual exomes allows for the detection of artifact with great sensitivity. The fact that we eliminated a number of now known artifacts and still observed a marked enrichment of RNS variation in these leukemia-associated genes is strong evidence that we are capturing biologically relevant phenomena. This enrichment was present in KMT2A-R+ IL. This provides more evidence that, while a powerful driver of leukemia, this rearrangement is insufficient to lead to a phenotype as severe as IL when acting in isolation. Rather, rearrangements may need the presence of a strong predisposition towards leukemia in order to bring it about *in utero*.

The notion that germline variation can strongly pre-dispose towards the development of cancer is becoming increasingly accepted[133,134]. Most studies, however, do not expand beyond known cancer pre-disposition genes. We, too, used prior knowledge as a proof of concept in this study.

However, we were, at the time of writing, unique in our expansion to a genome-wide interrogation of cancer-predisposing germline variation. We were able to do this because of a much increased sample size, and strong controls both performed in house and publicly available. Our genome wide study provided several new candidate genes, each of which might play a role in infant leukemogenesis. These genes might act independently and contribute to susceptibility to IL. Interestingly, though, many of the genes perform similar cellular functions. Specifically, ubiquitin ligation, transcriptional modifiers and RNA binding proteins are quite common in this group. This finding suggests that these processes might be particularly important in IL. It is also noteworthy that aside from KMT2B, KMT2C and PTEN, none of these genes have previously been implicated in either adult leukemia or a pan-cancer cohort[18]. There are probably several reasons for this. First, IL is a unique clinical entity[135,136], and, as such, should have unique genetic basis. Second, the fact that these genes are not frequently mutated in adult cancers supports the argument that the developmental context of genetic events might be as important as the events themselves. An insult that leads to cancer when it occurs early in development might be benign, or relatively so, in mature cell types and organisms. Finally, it is unlikely that a germline variant is equivalent to a somatic mutation. The subtle changes in activity or gene dosage that result from germline variation, but are present though the entirety of an individual's development and life, will likely have a different effect than the sudden loss or change in a gene much later in life as a result of somatic mutation.

Regardless of the explanation for this observation, it has implications for the management and treatment of IL. Currently, outcomes for these patients are poor. Early treatment regimens were insufficient and saw frequent relapse, but attempts to intensify treatments resulted in significant treatment-related morbidity and mortality[137,138]. Similarly, stem cell transplant is of limited

benefit for these patients[28]. The genes that we have shown to be highly variant in IL will provide insights into the molecular basis of IL. Since these genes were not seen in adult cancers, they might represent unique targets that will benefit IL patients where adult and childhood based treatment regimens have largely proved ineffectual. Further, taking into account the developmental timing and the differences between germline and somatic genetic insults will allow for a better informed therapeutic strategy.

In addition to informing our understanding of IL and potential treatments, this list of highly variant genes provides insights into the pathways and cellular functions that are critical for normal hematopoietic development. It seems that most of the genes work in some way to modulate or enforce a transcriptional program. This is clearly central to cellular function and behavior. It is likely that many of these genes, then, will be critical in several cell types. Still, the fact that they are observed to have such an increase in deleterious variation in IL specifically, suggests that they might be particularly important in the context developmental hematopoiesis. This system requires balance between self-renewal and differentiation[139], and perturbations during its establishment might be particularly damaging. Future studies into the specific functions of these genes in a hematopoietic context should be performed and will likely inform both development and IL alike.

Figure 3.1: Frequency of Rare Alleles in ExAC and EVS. The number of RNS alleles present in a given gene divided by the number of alleles sequenced in the cohort is plotted. It is evident from the plot that, despite very different sample sizes, this measure of variation is very similar between different cohorts, suggesting that the rates of genetic variation in a given gene are conserved.

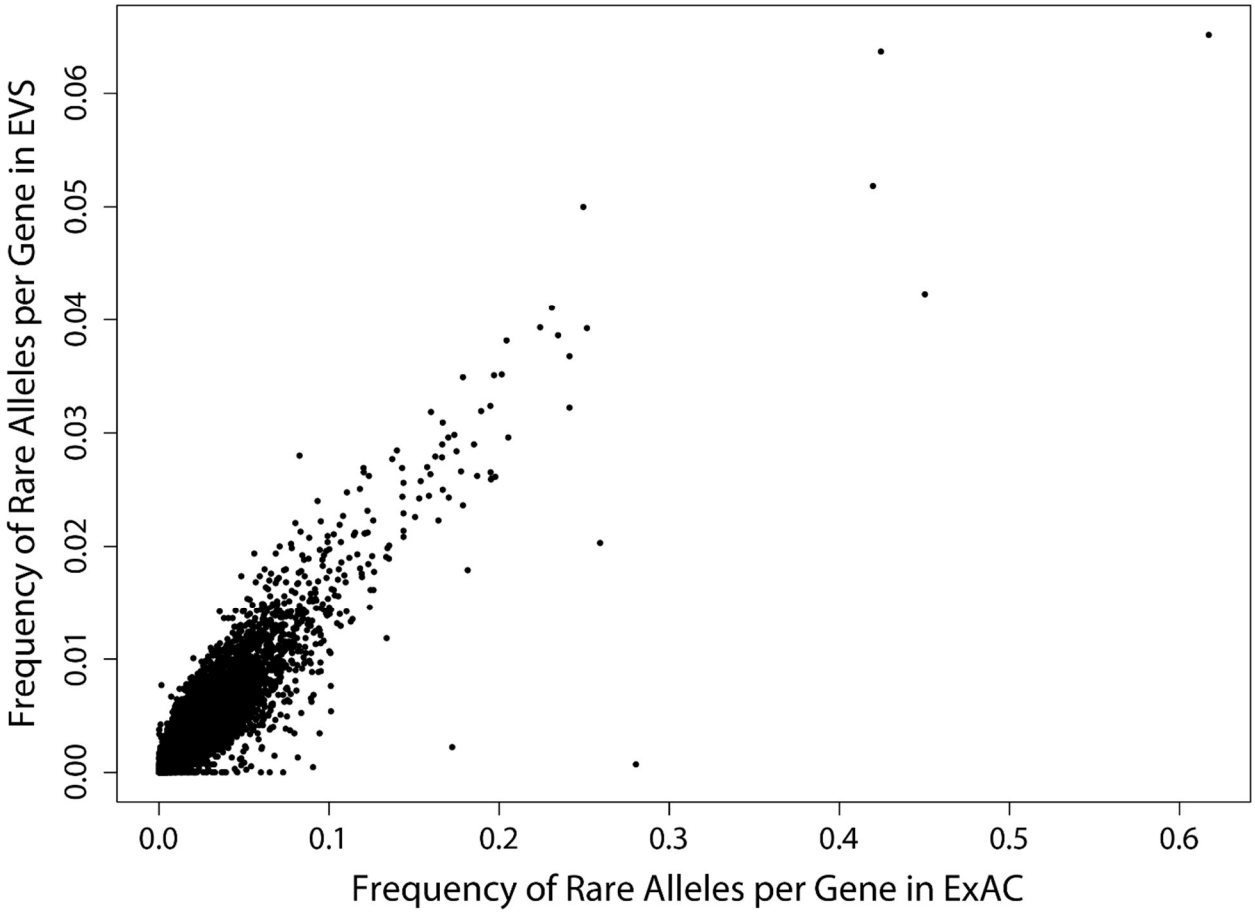


Figure 3.2: Frequency of Rare Alleles in ExAC and NHS. The same measure as Figure 3.1, but with samples from the NHS cohort replacing the EVS server on the y-axis. The linear relationship is still largely preserved, but, due to the much smaller sample size, there are a number of random fluctuations. To mitigate the influence of these random outliers, we applied a LoF intolerance test. Genes that are tolerant of LoF mutation are in gray; those that are not remain in black. KMT2B and KMT2C are depicted in red.

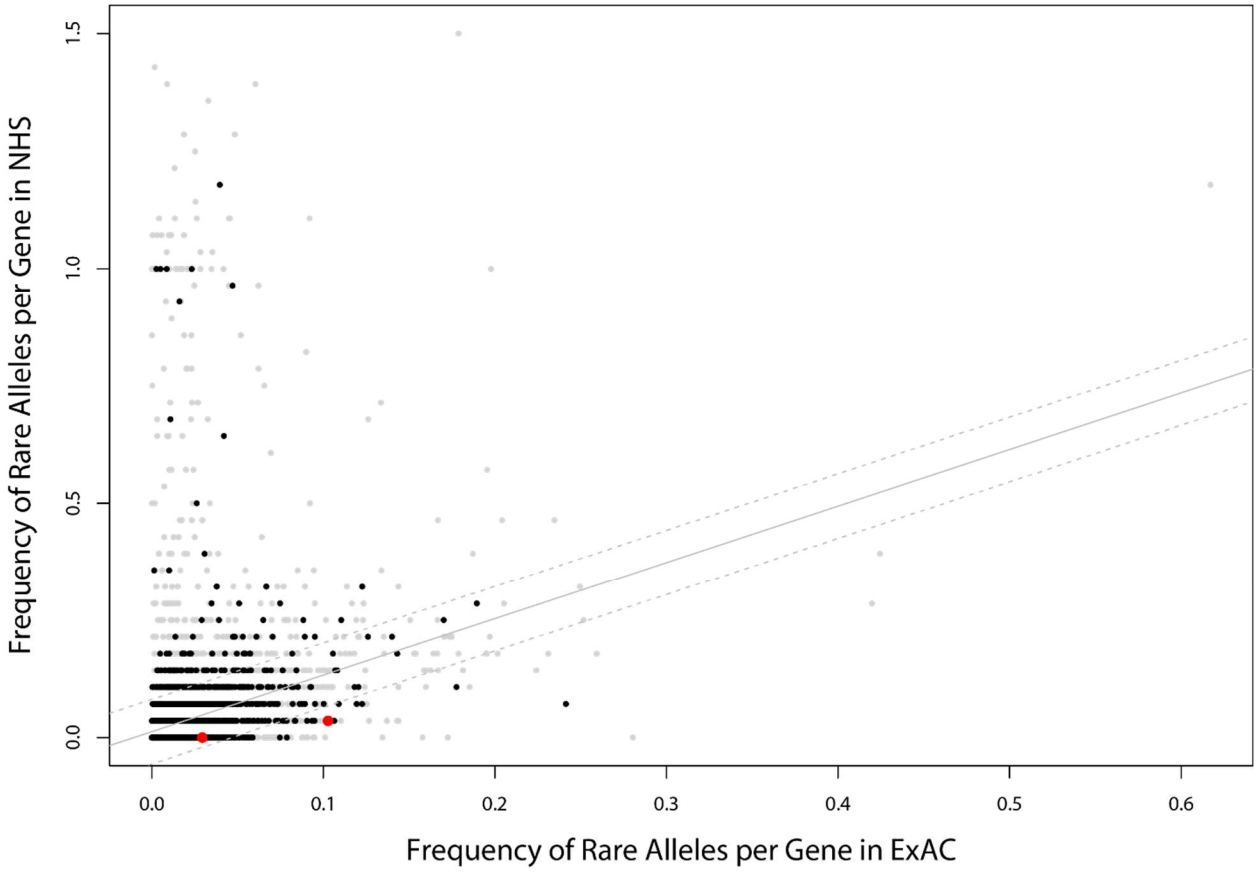


Figure 3.3: Frequency of Rare Alleles in ExAC and IL. The same measure as Figures 3.1 and 3.2, and colored as in 3.2. There is still a number of randomly "enriched" genes as an artifact of small sample size. However, relative to the NHS controls there are many more intolerant of LoF mutation. The solid gray line is a linear model fit to the data and the dotted gray lines are 5 standard deviations outside of this line.

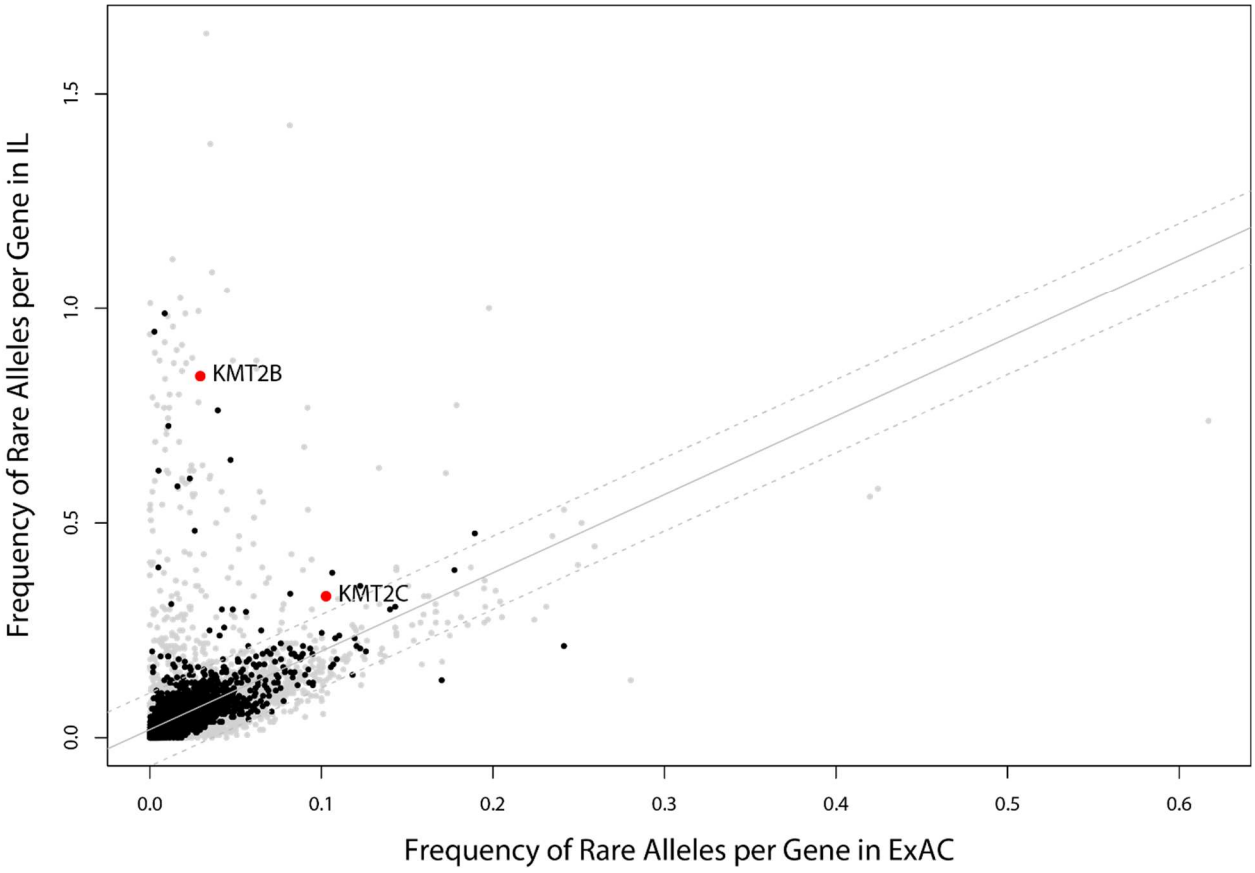


Figure 3.4: Frequency of RNS variants in KMT2 genes and COMPASS complexes. The number of RNS variants per individual is plotted for each of the KMT2 family genes along with the same measure for genes that are either shared between both COMPASS complexes or unique to the KMT2A/B complex of the KMT2C/D complex. Complex members are defined here[132]

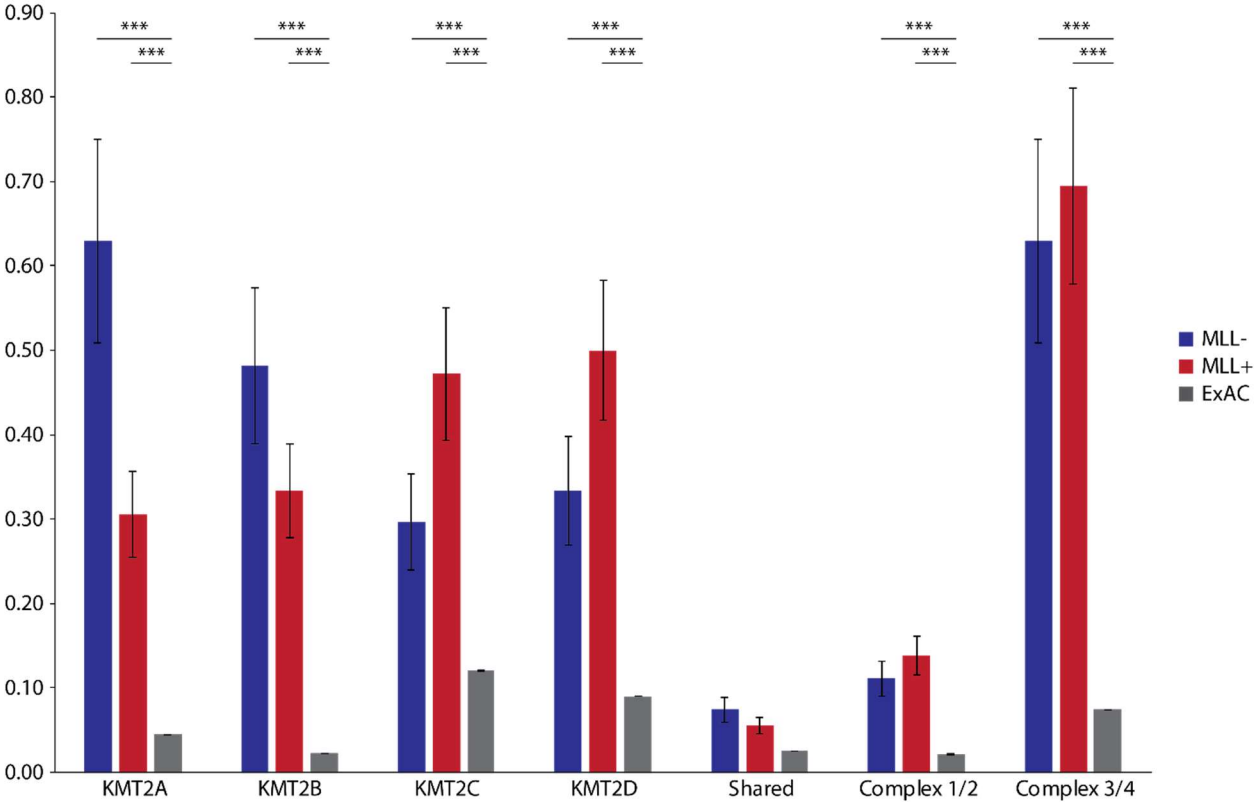


Figure 3.5: Location of variants in KMT2 genes in IL cases and controls. Lollipop plots for KMT2A (MLL1), KMT2B (MLL2), KMT2C (MLL3) and KMT2D (MLL4). While there does not seem to be variant hotspots, there is a clear difference in the number and types of variants present in the IL cases relative to controls.

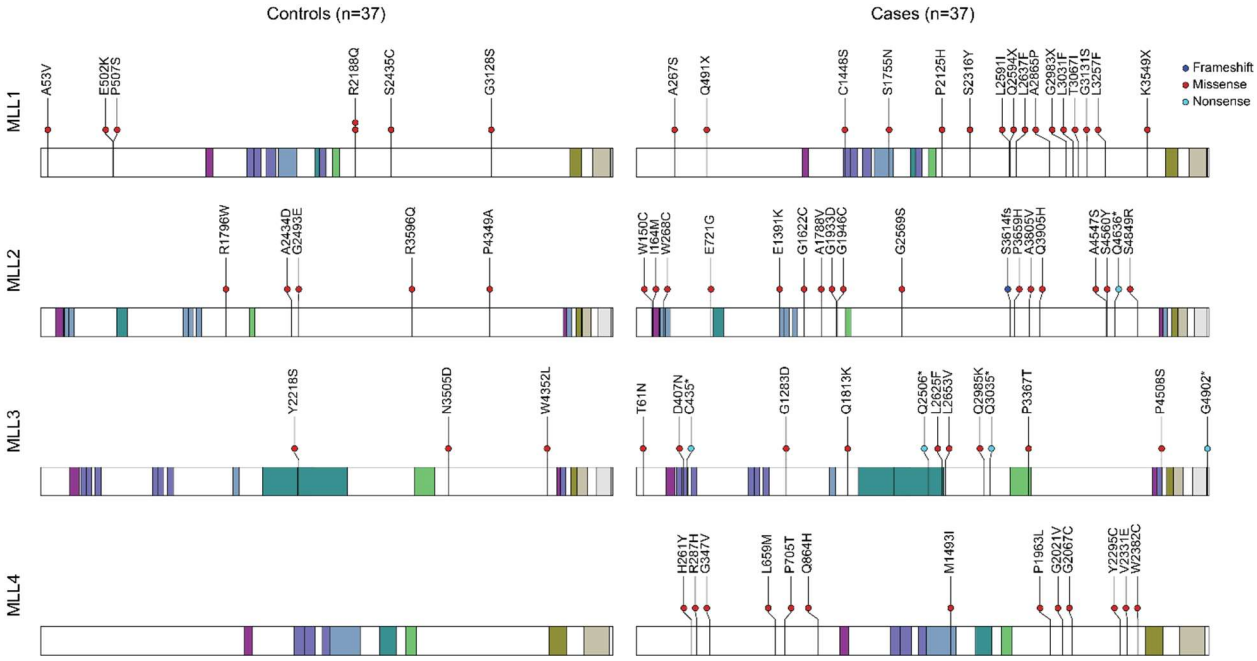


Figure 3.6: Co-occuring variants by gene and patient. The variants present in each patient in each of the top variable genes was plotted.

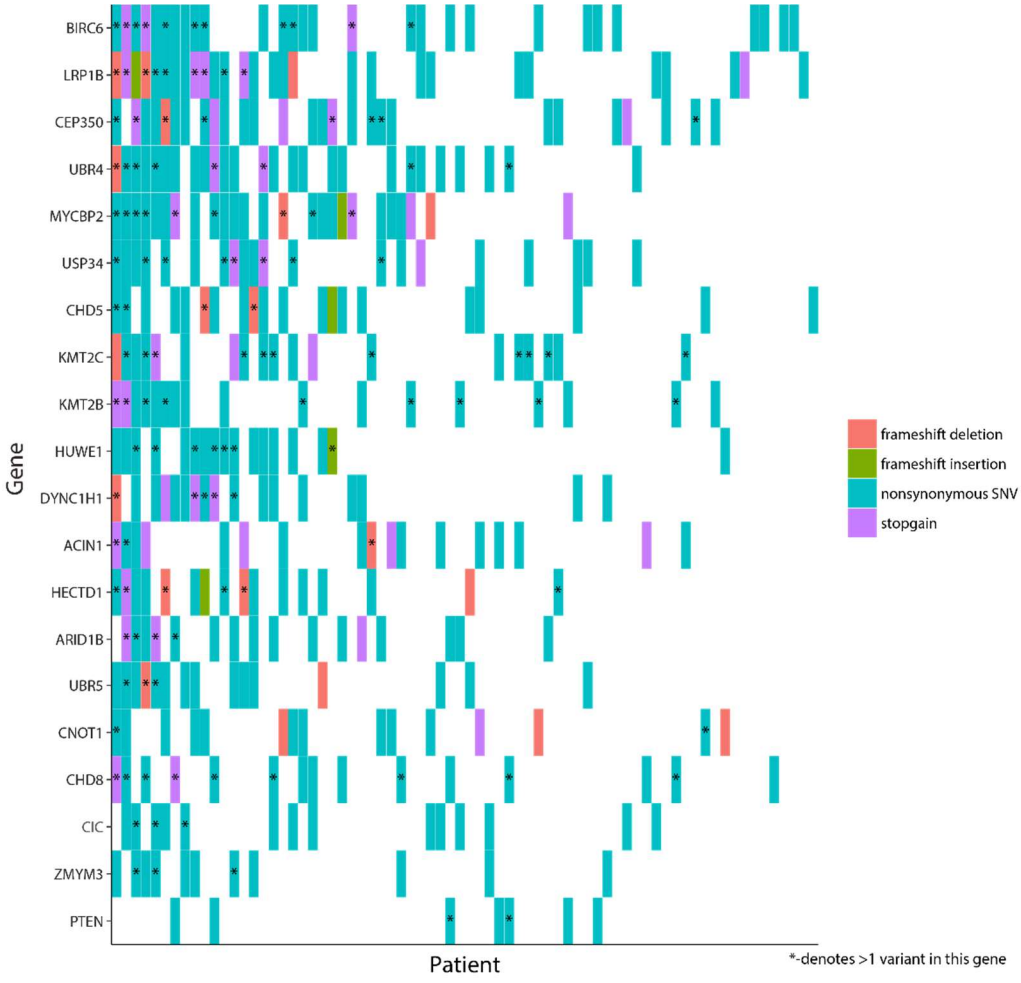


Table 3.1: Enrichment of RNS Variants in Leukemia-associated Genes.

Leukemia Subtype	KMT2A-R status	ALL Enrichment p-value	AML Enrichment p-value
ALL	+	0.001	-
ALL	-	0.002	-
AML	+	-	1.25e-10
AML	-	-	4.67e-14
Control	-	NS	NS

Table 3.2: Genes with significantly more RNS variation in IL compared to ExAC. This list was further filtered by LoF intolerance and the control NHS sample.

KMT2C	MYCBP2	CIC
KMT2B	DYNC1H1	UBR5
PTEN	CEP350	HECTD1
CHD5	CNOT1	USP34
CHD8	HUWE1	BIRC6
LRP1B	ACIN1	UBR4
ZMYM3	ARID1B	

4. Modeling Hematopoietic Development *in vitro*

4.1 Introduction

The data from our germline exome sequencing yielded some important insights into the biology of IL. The notion that germline variants play a role in the development of IL, which was not widely anticipated, seems quite evident in our data. This finding and the *in utero* onset of IL suggest that this disease is, at least partially, the result of a defect in the establishment of a normal hematopoietic system. While it is not feasible to test this directly in patients, the list of genes that are more highly variant in IL patients provides several potential candidates for *in vitro* modeling. The evidence that each of them might have a role in infant leukemogenesis is strong, however, we focused on KMT2C as an initial proof of concept for several reasons. The role of KMT2A rearrangements in IL is clear[22,31,35,140,141]. The family members KMT2B and KMT2C are significantly mutated in a pan-cancer dataset[131]. KMT2C is also present in our original list of leukemia associated genes. Finally, after our analysis of the KMT2A-R- IL patients, KMT2C appeared to have compound heterozygous variants in every case of infant AML and most cases of infant ALL. Our subsequent analysis tempered this finding, but the focus on KMT2C is still supported in these later data as well.

To appropriately model developmental hematopoiesis, and how it might go awry in the development of IL, we turn to an hPSC-based directed differentiation system[142]. There are numerous advantages to using this system. The cells are human and thus should accurately reflect the role of human genes. Much of hematopoiesis is conserved in vertebrates, but the role of certain genes is not always consistent across species[143]. The timeframe for these experiments

is shorter than that required to perform animal studies. The two programs of hematopoiesis that occur during embryonic development can be separated and explored individually[144]. A quick and robust method to introduce transgenes that are expressed throughout the differentiation process has already been developed[145]. Finally, a great deal of work has already been done to establish and characterize a number of time points throughout the differentiation process, allowing for the rapid and relatively easy interrogation of a number of developmental stages along the hematopoietic axis[56,93,144].

4.2 Materials and Methods

4.2.1 iPSC lines

The control line (IC1) used in this study was generated from fibroblasts taken from a healthy adult male. They were reprogrammed using the CytoTune-iPS 2.0 Sendai Reprogramming Kit from Invitrogen (Cat. #A16517) according to the manufacturer's instructions. Briefly, this kit uses a non-inserting Sendai Virus expressing the four reprogramming factors, OCT4, SOX2, KLF4 and MYC (OSKM).

4.2.2 CRISPR knockout of KMT2C

Guide RNAs targeting exon 3 of KMT2C were generated and checked for off-target effects. IC1 cells were transfected with Cas9 and the validated guide RNAs. The resulting cleavage sites were rejoined by non-homologous end joining, resulting the expected deletion events. Screening revealed that several knock out clones were generated (Figure 4.1). The clone used in this work is dubbed IC1_MLL3KO

4.2.3 hPSC growth and maintenance

The day before hPSCs are to be split or thawed, seed an appropriate number wells with irradiated MEFs. Aspirate the media from each well to be split. Add 1 mL room-temperature trypsin-EDTA to each well. Incubate at room temperature for 1 minute. Aspirate the trypsin and stop the reaction by adding 1 mL stop media per well. Gently scrape the cells until they have all lifted off the plate. Add 1 mL wash media to each well, and triturate 3-5 times with a 2 mL serological pipette. This should result in relatively uniform clusters of 5-15 cells. Harvest each well and centrifuge at 1200 rpm for 5 minutes. Resuspend the cells in 2 mL hESC media per well to be seeded and aliquot the cells onto MEFs. Typically, the cells will be passaged at a 1:12 split every 5 days.

4.2.4 Differentiation Protocol

Day 0: Generation of Embryoid Bodies (EBs)

The day before EB generation, cells are split and plated onto matrigel-coated dishes to deplete the MEFs. The MEF-depleted cells are treated with 1 mL of a 1:4 solution of Trypsin-EDTA:PBS for 30s. The trypsin is removed and the reaction is stopped by the addition of 1 mL stop media per well. The scrape each well with a cell-scraper until all cells have lifted from the plate. Add 1 mL of wash media to each well and triturate with a 2 mL serological pipette until the cells have formed clusters of 10-20 cells. Collect each well into a 50 mL conical tube and centrifuge at 800 rpm for 5 minutes. Aspirate the supernatant and centrifuge for an additional 5 minutes. Resuspend the cells in 2 mL of Day 0 media for every 2 wells of cells harvested and distribute into polyheme-coated 6-well plates. Incubate at 37C 5% O2.

Day 1: Feeding EBs

Add 2 mL of Day 0 media supplemented with 10ng/mL bFGF to each well of the culture

Day 2: Induction 1

This step should be performed 42 hours after EB generation. Harvest up to 12 wells of culture into one 50 mL conical and centrifuge at 400 rpm for 5 minutes. Aspirate the supernatant and resuspend in pure IMDM. Spin another 5 minutes at 400 rpm. This step will remove much of the debris that is present in the cultures at this point. After the second centrifugation, aspirate the supernatant and resuspend 2 mL in Day 2 media for each well of culture harvested. Return the aggregates to the original polyheme-coated plates. Incubate at 37C 5% O2.

Day 3: Dissociation for Mesoderm Analysis or Induction 2

If a mesoderm analysis will be performed, harvest the cells into a 50 mL conical and centrifuge at 1200 rpm for 5 minutes. Resuspend in 3 mL trypsin-edta. Incubate in a 37C water bath for 5 minutes. Add 3 mL of stop media and centrifuge at 1200 rpm for 5 minutes. Resuspend the cells in IMDM supplemented with 2% FCS and pass through a filter to ensure a single cell suspension. This suspension is ready for downstream analysis.

If no mesoderm analysis will be performed, harvest the cells into a 50 mL conical and centrifuge at 1200 rpm for 5 minutes. Wash once in IMDM and centrifuge again. Prepare 2 mL of Day 3 media for each well of culture. Resuspend the washed aggregates in Day 3 media and return to the same polyheme coated plates. Incubate at 37C, 5% O2.

Day 6: Feeding EBs and supplementing cytokines

Prepare 2 mL of Day 6 media per well of culture and add it to each well.

Day 8+: Dissociation

Harvest up to 12 wells of culture into a single 50 mL conical. Centrifuge at 1200 rpm for 5 minutes. Aspirate the supernatant and resuspend the cells in 3-6 mL trypsin-edta. Incubate for 8 minutes in a 37C water bath. Stop the reaction by adding 3-6 mL stop media and centrifuge at 1200 rpm for 5 minutes. Aspirate the supernatant and resuspend in 3-6 mL Collagenase Type II. Incubate in a 37C water bath for 30-60 minutes. Add 3-6 mL stop media and centrifuge at 1200 rpm for 5 minutes. Aspirate the supernatant and resuspend the cells in IMDM supplemented with 2% FCS. Pass the cells through a filter to ensure a single cell suspension. The cells are ready for downstream assays and/or analysis.

4.2.5 Flow Cytometry and antibodies

Table contains a list of the antibodies, fluorophores and concentrations used for each experiment.

Once the cells are in a single cell suspension, wash them 5 times (day 3) or 2 times (all other days) in IMDM+2%FCS. Resuspend the cells at a concentration of up to 1 x 10^6 cells/100 uL. Add the appropriate antibodies and incubate on ice for 30 minutes. Add 1 mL IMDM+2% FCS and centrifuge at 1200 rpm for 5 minutes. Aspirate the supernatant and resuspend the cells at a concentration of 5x10^6 cells/mL.

4.2.6 Endothelial to Hematopoietic Transition (EHT) Assay

Resuspend CD34+, CD43- endothelial cells, or CD34+, CD43-, CD73-, CD184- hemogenic endothelial cells in Day 8 media at a concentration of 300,000 cells/mL. Add 30 uL of cell suspension per well of a matrigel-coated 24-well plate and incubate overnight at 37C 5% O2.

The next day, add 1 mL Day 8 media to each well. The cells will undergo EHT over the next several days. They can be harvested at Day 8+8 for definitive (CHIR99021-conditioned) cultures or at Day 9+8 for primitive (IWP2-conditioned) cultures.

4.2.7 Serial Replating Assay

Harvest cells that have undergone EHT and centrifuge them at 1200 rpm for 5 minutes. Aspirate the supernatant and resuspend them in 1 mL Day 8 media. Count the cells with a hemocytometer and add enough Day 8 media to arrive at a concentration of 250,000 cells/mL. Add 1 mL of cell suspension to each well of a non-adherent 24 well plate. Repeat this process every 7 days until the cultures have either failed or demonstrated serial replating capacity.

4.2.8 Methylcellulose-based Colony Forming Assay

Add 50,000 cells from a single cell suspension to a 2.5 mL aliquot of MethoCult (StemCell Technologies cat. #04034). Vortex thoroughly. After the mixture has settled and the air bubbles have largely disappeared, draw up 1 mL of Methocult into a syringe and aliquot into a 35-mm dish. Repeat for the remainder of the Methocult and cell mixture. Place all 35-mm dishes into a 150-mm with a single uncovered 35-mm dish filled with water. Place the larger dish into the 37C incubator. Colonies are counted 7-9 days after plating.

4.2.8 T-cell Assays

Resuspend cells in T-cell Media and plate onto OP9-DL4 stroma. Passage the cells every 4 days, splitting them 1:5. After 17-21 days in culture, proceed with flow cytometry.

4.2.9 StemPro

Thaw StemPro34 supplement overnight at 4C. Add supplement to 500 mL StemPro34 Medium. Add 5 mL Penn/Strep. Mix by gently inverting. Incubate at 37C for 30 min. Aliquot and store at 4C.

4.2.10 AAVS Targeted MLL-AF9 Construct

The sequence of a common MLL-AF9 fusion was obtained from the NCBI. This construct was synthesized and cloned into the TRE AAVS plasmid by Genewiz.

4.2.11 AAVS Transfection

Treat a 5-day old culture of hPSCs with trypsin until they are a single cell suspension. Count the cells and seed 250K/well of drug-resistant MEFs. Include ROCK inhibitor in the hPSC medium. The next day, change the medium to P/S-free hPSC medium with no ROCK inhibitor. Mix 1.2 ug neo-rtTA plasmid, 1.2 ug TRE-MA9 and 0.3 ug of each ZFN plasmid in 100 uL of P/S-free IMDM. Add 9 uL XtremeGENE 9 and incubate 25 minutes at room temperature. Slowly add 100 mL DNA mixture to each well. The next morning, feed with P/S-free hPSC medium. Add 2.5 ug/mL puromycin and incubate 2.5 days. Return the cells to P/S-free medium for 1 day. Incubate for the remaining days of selection in medium containing 20 ug/mL G418. Manually pick Individual clones that have grown out for genotyping.

4.2.12 AAVS Screening

For the successful generation of a dox-inducible cell line, both wildtype alleles must be lost, and one each of the rtTA and the TRE plasmids must be in their place. Thus, 3 PCR reactions are performed. First, to check for wildtype allele loss, second for the rtTA integration and third for the TRE integration. The primers used are:

AAVS WT HET F: CCC CTA TGT CCA CTT CAG GA

AAVS WT HET R: CAG CTC AGG TTC TGG GAG AG

CAG AAV scrn FWD: TCCTGGGCAAACAGCATAA

TRE AAV scrn REV: GAAGGATGCAGGACGAGAAA

TRE AAV scrn FWD: GCAATAGCATCACAAATTTCAC

TRE AAV scrn REV: GAAGGATGCAGGACGAGAAA

4.3 Results

As there is known heterogeneity in the behavior of different pluripotent cell lines, we first sought to establish the baseline differentiation characteristics in IC1, our control cell line derived from a healthy adult male. To start, we used the same differentiation schedule and media formulations as has been previously reported[142] (Figure 4.2). As expected, IC1 cells differentiated for 3 days in the presence of BMP4 and bFGF acquired a KDR+ mesoderm phenotype (Figure 4.3). This mesoderm expressed either high or low-absent levels of CD235a when treated with the small molecules IWP2 or CHIR90221 respectively. These markers are indicative of mesoderm that is primed to undergo either primitive or definitive hematopoiesis. We continued these cultures, changing the media and adding appropriate cytokines until day 8 of differentiation. Again, this line exhibited the expected phenotypes. Specifically, cells that were treated with IWP2 had a CD34+, CD43- population that was slowly maturing into a CD34-, CD43+ population with a gradient of cells at various points in this transition, indicative of active and ongoing primitive hematopoiesis (Figure 4.4). To further show establish the efficacy of primitive hematopoiesis,

we performed a colony forming assay and observed robust, multi-lineage colony forming potential (Figure 4.5). By contrast, the cells that were treated with CHIR90221, and were thus primed exclusively for definitive hematopoiesis, had only a CD34+, CD43- population. This population requires additional signaling to undergo hematopoiesis, so to further characterize the behavior of this line, we sorted these mixed endothelial cells and used them as inputs for both an endothelial-to-hematopoietic transition (EHT) assay, and a T-cell assay.

In the EHT assay, sorted CD34+,CD43- cells were plated onto matrigel coated plates and allowed to adhere in low volume overnight. Cytokine rich medium was added the next day and the cells were observed for 8 days. Over this time, the cells first divide and expand to form an endothelial-like monolayer of adherent, close packed cells. Various cells begin to become more round and protrude up from their surrounding endothelium. Eventually, these cells will release completely to enter suspension and take on a round, small, bright phenotype consistent with a hematopoietic cell. Figure 4.6 shows a series of micrographs detailing this process. After 8 days, the hematopoietic cells can be re-plated and followed over several weeks. Initially, the culture is predominantly comprised of hematopoietic progenitors that have the ability to divide several times, and differentiate into a variety of mature hematopoietic cell types. Over time, these cells become terminally differentiated and lose the ability to self-renew. As a result, the number cells in culture initially expands, but eventually contracts as the cells senesce and die (Figure 4.7).

The T-cell assay takes the same cell population as input. The cells are plated onto a layer of OP9-DL4 stroma that provides high levels of Notch signaling requisite for T-cell maturation. They are passaged regularly over 21 days and then interrogated by flow cytometry. As expected,

the differentiated IC1 cells were able to differentiate into T-cells as evidenced by the expression of both CD4 and CD8 (Figure 4.8).

Together, these experiments establish that IC1 cells behave as expected in this differentiation protocol. They give rise to both primitive and definitive hematopoietic cells that form colonies and persist in culture for an expected timeframe.

We next turned our attention to the KMT2C KO line that we derived from our IC1 cells. We again sought to establish the ability of these KMT2C KO cells to undergo hematopoietic differentiation and determine if any differences existed between the IC1 and KMT2C KO cell lines during the differentiation process. We proceeded as before, first examining the cultures after 3 days. As shown in Figure 4.9, the cells had taken on the expected mesodermal phenotype, with appropriate responses to induction in the presence of either Wnt stimulation or block. Similarly, after 8 days in culture, the cells had acquired the anticipated endothelial phenotype (Figure 4.10). However, when we performed a colony forming assay on primitive cells we observed a marked decrease in the number of colonies that arose (Figure 4.5). More strikingly, when we performed EHT and T-cell assays, we observed a complete lack of definitive hematopoiesis (Figures 4.8 and 4.11).

To further explore this result, we repeated the differentiation, this time adding CD73 and CD184 antibodies to the flow panel on day 8. CD73 and CD184 enable the identification of the different types of endothelium present in the CD34+ CD43- definitive cells[56]. When we added these markers, the reason for the KMT2C KO cell line's failure to undergo EHT and make T-cells was readily apparent. The endothelial cells that have hematopoietic potential, hemogenic endothelium (HE), are CD73- CD184-, this population is present in the IC1 cultures but effectively absent in

the KMT2C KO cultures (Figure 4.12). To ensure that the lack of HE as defined cytometrically, we sorted all endothelial cells (i.e. CD34+,,CD43-) from both IC1 and KMT2C KO cultures, and performed an EHT assay. As before, only the IC1 cultures gave rise to any meaningful amount of cells in suspension. Still, we collected all of the cells from each line after 8 days of EHT and performed a colony forming assay. The IC1 line gave rise to multiple robust colonies, while the KMT2C KO produced effectively none (Figure 4.13). A similar, though less severe, phenotype was seen in primitive differentiations as well (Figure 4.5). This indicates that KMT2C is required for efficient establishment of hematopoiesis in this *in vitro* model.

4.4 Discussion

hPSC cells that lack KMT2C are impaired in their ability to undergo primitive hematopoietic differentiation in our system. Similarly, KMT2C knockout cells fail to specify definitive HE in this same system. The observation of this striking phenotype validates our bioinformatic analysis of IL exomes. We found that this gene much more highly variable in IL patients than population controls. We postulated that these variants resulted in impaired function of KMT2C and thereby led to aberrant hematopoietic development. In our hPSC based model of hematopoietic development, we see marked dysfunction in the absence of KMT2C. It will be interesting to see if other genes that were flagged in our analysis will have similarly strong phenotypes in this or other systems.

While this study works well as a proof of concept, there are several caveats and open questions. First, the complete lack of definitive HE and impaired primitive hematopoiesis do not immediately suggest leukemia. There are several possible explanations for this. First, we have introduced only one genetic lesion when our bioinformatics analysis, as well as years of cancer

genomics, suggest that dysfunction in several genes is required for frank tumorigenesis[17]. Also, the infants in our sequencing studies generally have heterozygous non-synonymous point mutations in KMT2C, whereas our cell line has a complete knockout of the gene. Thus, our cells should have a complete lack of KMT2C function, but the IL patients are much more likely to have a hypomorphic allele. This difference might result in a less severe phenotype than the one we observed. Either way, a block in differentiation is consistent with cancer development[146–148], especially when combined with a proliferative insult from KMT2A rearrangements or RAS mutations that are relatively common in IL[22,149].

KMT2C disruption led to dysfunction in both the primitive and definitive hematopoietic programs. This emphasizes the importance of this gene in these contexts. However, this finding does not provide any insight into the cell-of-origin in IL. Since both programs were affected in the absence of this gene, the possibility still remains that either program could be the one that is hijacked in infant leukemogenesis. The possibility that aberrant primitive hematopoiesis might be transformed into leukemia is compelling. Given the clinical and phenotypic differences between IL, later childhood leukemia, and even other KMT2A-R+ leukemia later in life, a primitive cell-of-origin in the former is quite possible[141,150]. Our results are consistent with this possibility, but are far from confirming it. However, several groups have shown that hPSCs derived from AML patients re-acquire leukemic properties when differentiated into hematopoietic cells[151,152]. Interestingly, the differentiation protocols used in these studies are likely to form a mixture of primitive and definitive cells. In this case, the primitive program, which is highly proliferative early on, is likely to represent a majority of the cells in these cultures. Thus, the cells that are able to serially re-plate and establish leukemia in transplanted mice, are very possibly derived from the primitive program. Establishing whether this is the case

is an important goal, and is actively being explored. This knowledge will definitely inform developmental biology and the etiology of IL. It could also lead to much needed improvements in IL treatment.

Pluripotent cells derived from patients with various hematopoietic disorders have stereotypic lesions that are presumably the drivers that lead to the re-acquisition of leukemic phenotypes when these cells are differentiated back into hematopoietic lineages[151,152]. These studies demonstrate that leukemia can be modeled by *in vitro* hPSC differentiation strategies. However, these cells start from cells that have already transformed into frank leukemia. Our attempts to model IL work in the opposite direction; starting from phenotypically normal cells, and adding genetic insults until transformation is achieved. The deletion of KMT2C was not able to recapitulate IL in our differentiation, but this single insult was expected to be insufficient for a complete leukemic transformation. With the addition of other insults, we anticipate that we too will be able to achieve a leukemia-like phenotype from our cells. This model will be more relevant to IL because it will be based on sequencing studies in IL, and will be able to distinguish the cell-of-origin relevant to IL. Further, we will have a closed system in which the genetic insults will all be known, and the cell line in which the insults were introduced is epigenetically normal as evidenced by its ability to undergo hematopoietic differentiation normally.

Figure 4.1: Schema of KMT2C KO generation. The CRISPR cut sites are shown and the various alleles that resulted from the non-homologous end joining are shown. Each introduces a frameshift mutation very early in the protein.

Figure 4.2: Differentiation schema. A cartoon depiction of the differentiation process that we used. Cells that are conditioned with CHIR09221 and become definitive, are in blue, while those in yellow were treated with IWP2 and will have a primitive phenotype.

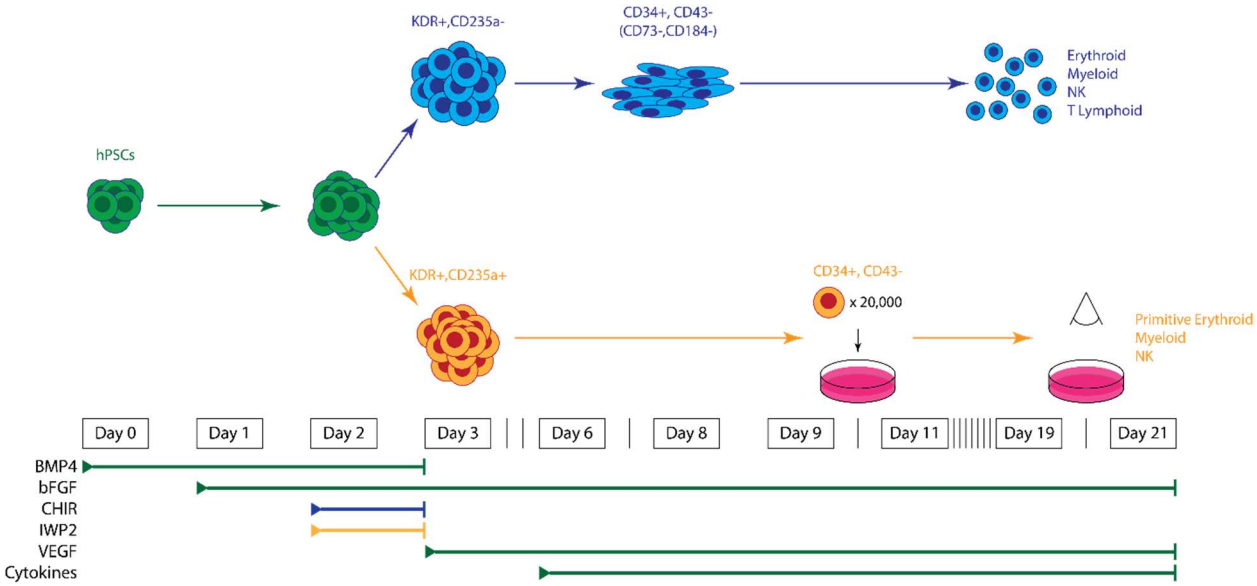
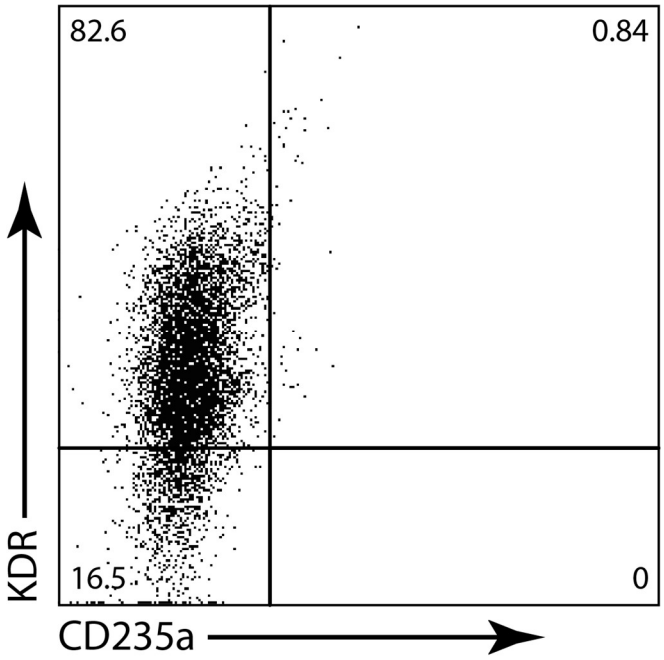


Figure 4.3: Day 3 flow cytometry in control cells. Panel A depicts a CHIR09221-treated culture. The lack of CD235a cells indicates that there is no primitive specified mesoderm present. Conversely, the culture shown in Panel B was treated with IWP2 to specify primitive primed mesoderm.

a)



b)

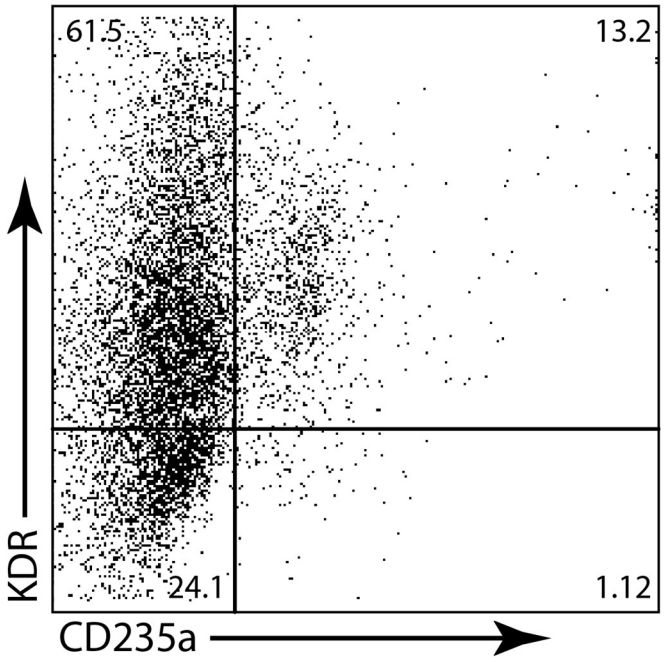


Figure 4.4: Day 8 flow cytometry in control cells. As in Figure 4.3, the control cells that were treated with the two modulators of Wnt-signaling have the expected phenotypes in flow cytometric assays. The CD43 expressing cells present in the IWP2 treated culture are indicative of active primitive hematopoiesis.

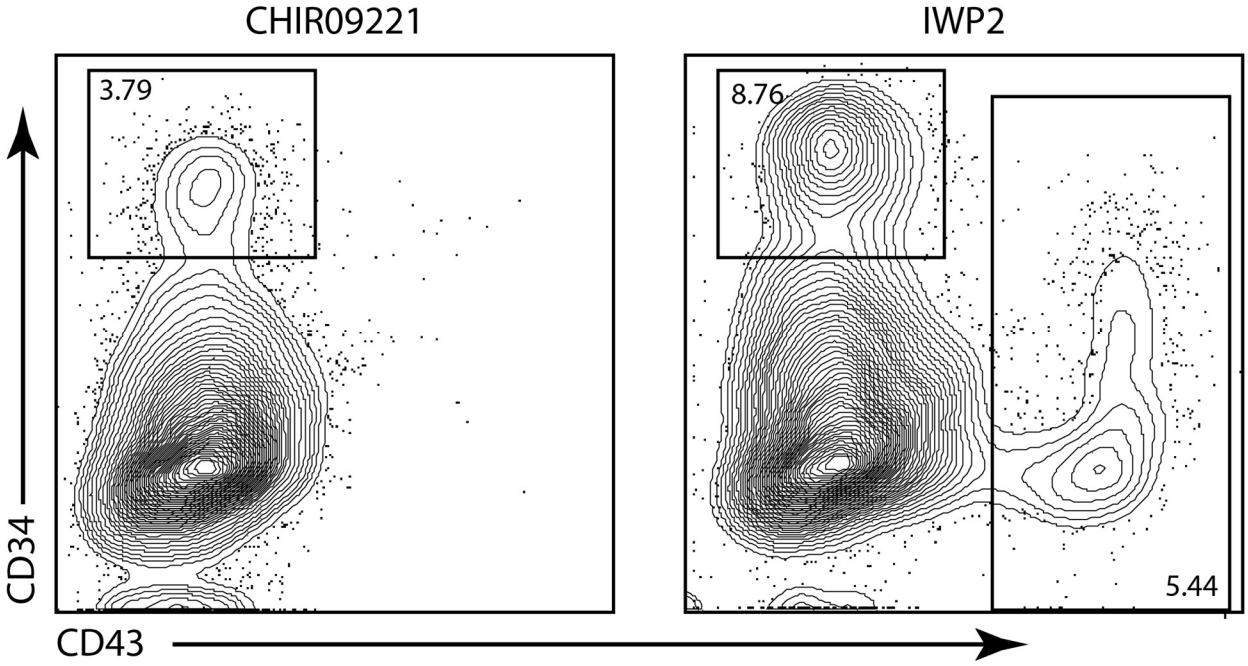


Figure 4.5: Colony-forming assays in primitive-specified cells. The average number of myeloid and erythroid colonies per 10,000 cells is depicted. There is a significant decrease in both lineages in KMT2C (MLL3) KO cells.

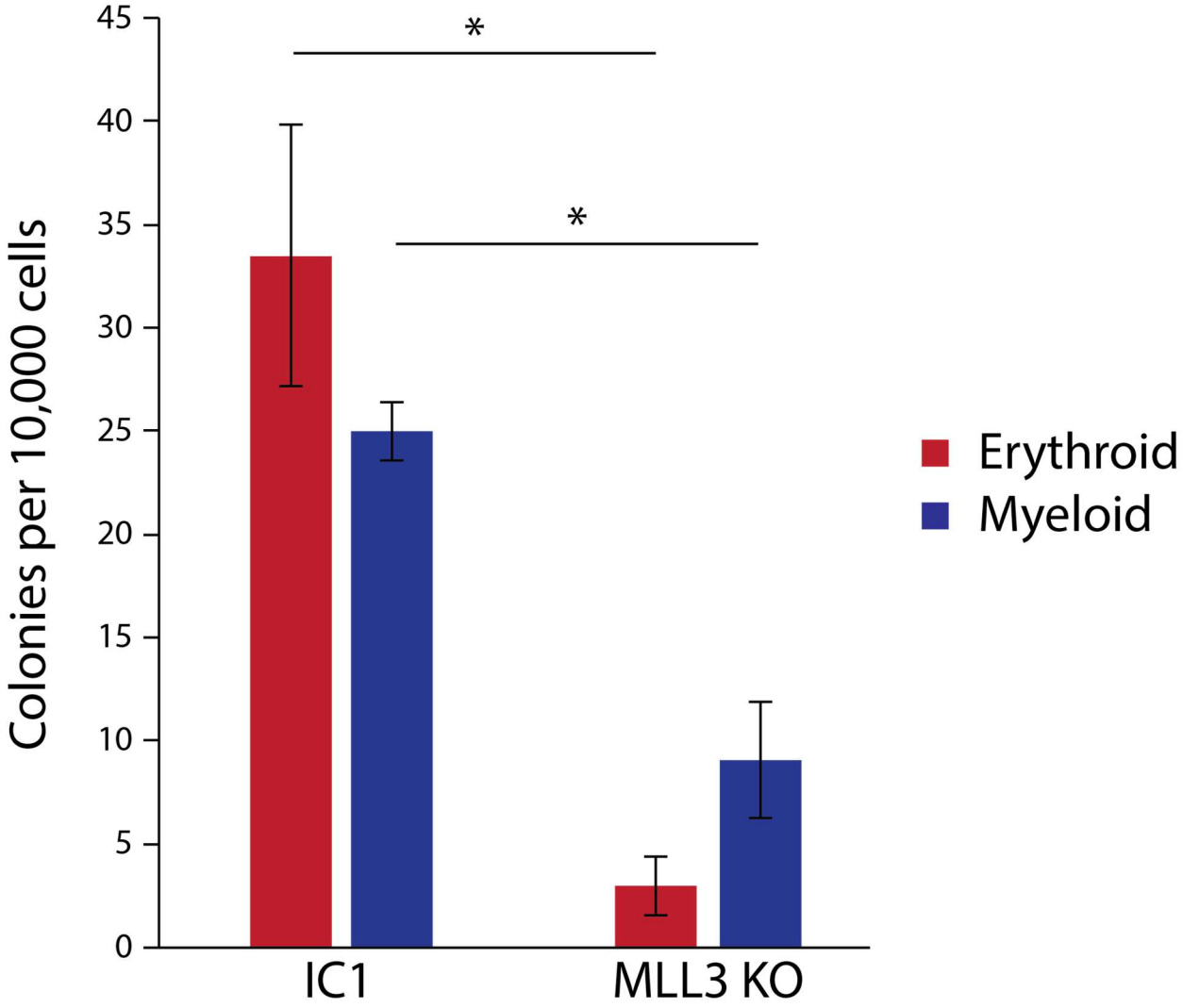


Figure 4.6: Micrographs of EHT in control cells. The EHT process occurs over several days and is documented below. The cells initially have an endothelial phenotype, and fill in to create a continuous monolayer. Over time, certain cells will become rounder and eventually non-adherent to release into the medium.

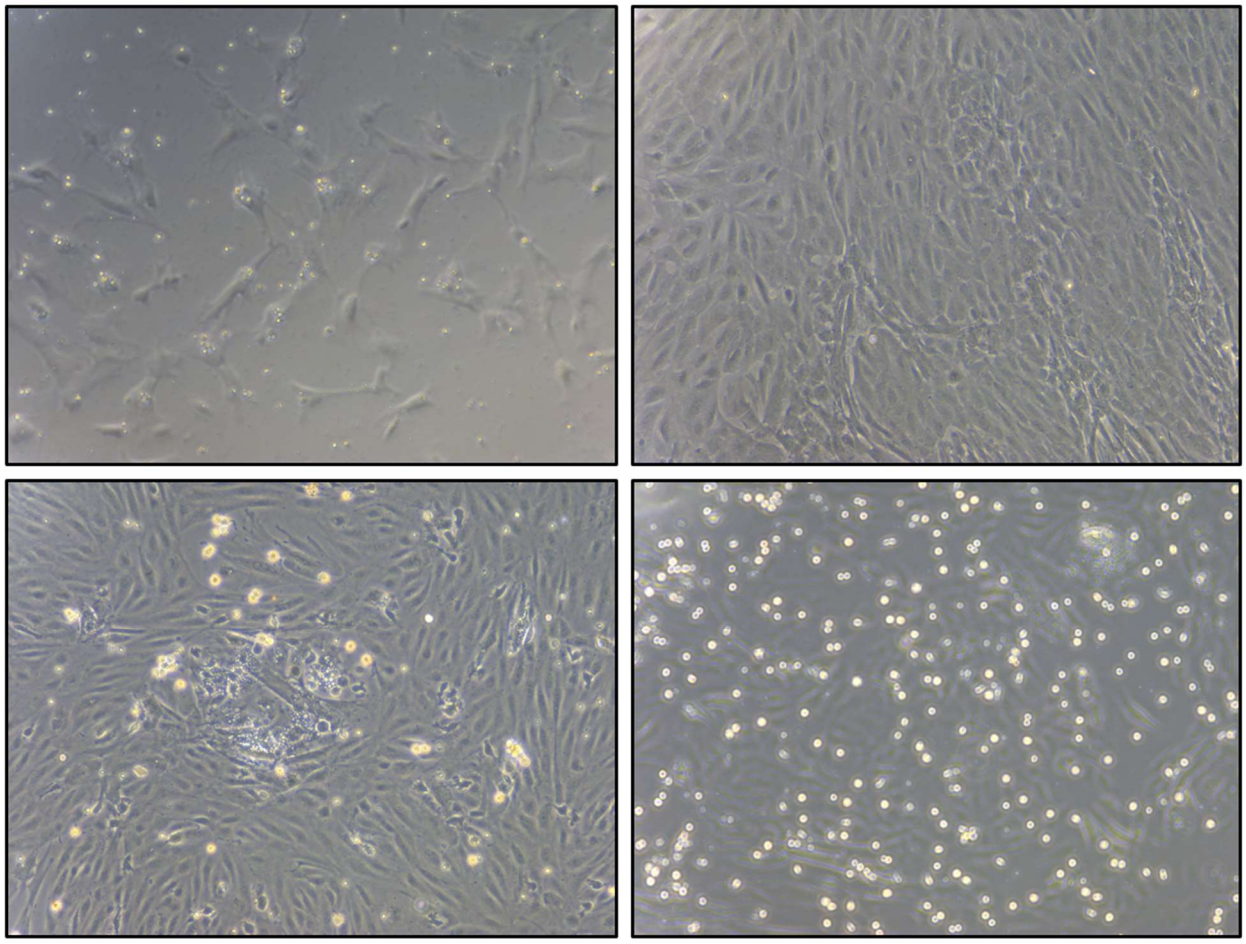


Figure 4.7: Expansion of serially passaged control cells. The cells that undergo EHT are a mix of progenitors and more mature cells. As a group they are capable of expanding over several days. However, HSCs are not supported in this assay, so eventually the ability to expand is exhausted and the culture collapses.

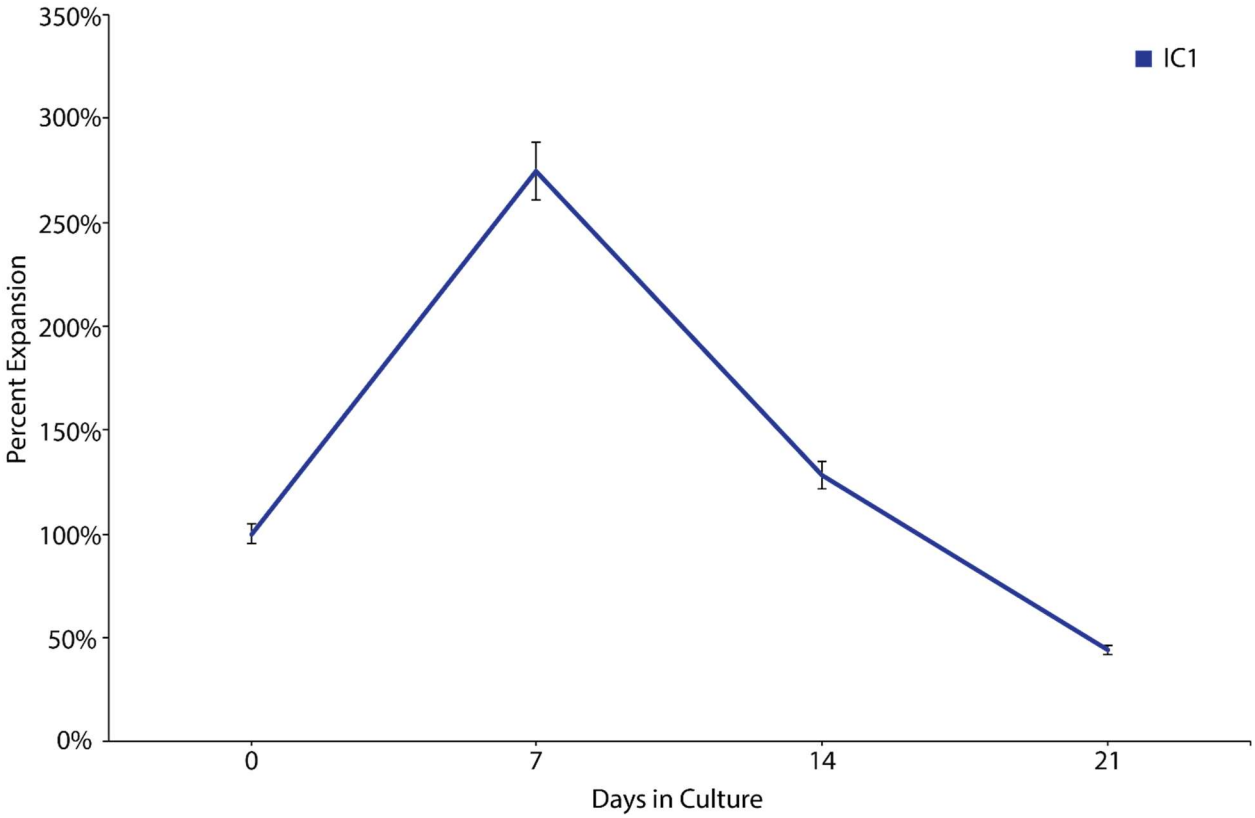


Figure 4.8: T-cell assay in control and KMT2C KO cells. Control (IC1) cells have robust T-cell potential indicating successful definitive hematopoiesis. KMT2C (MLL3) KO cells, conversely, repeatedly failed to generate T-cells.

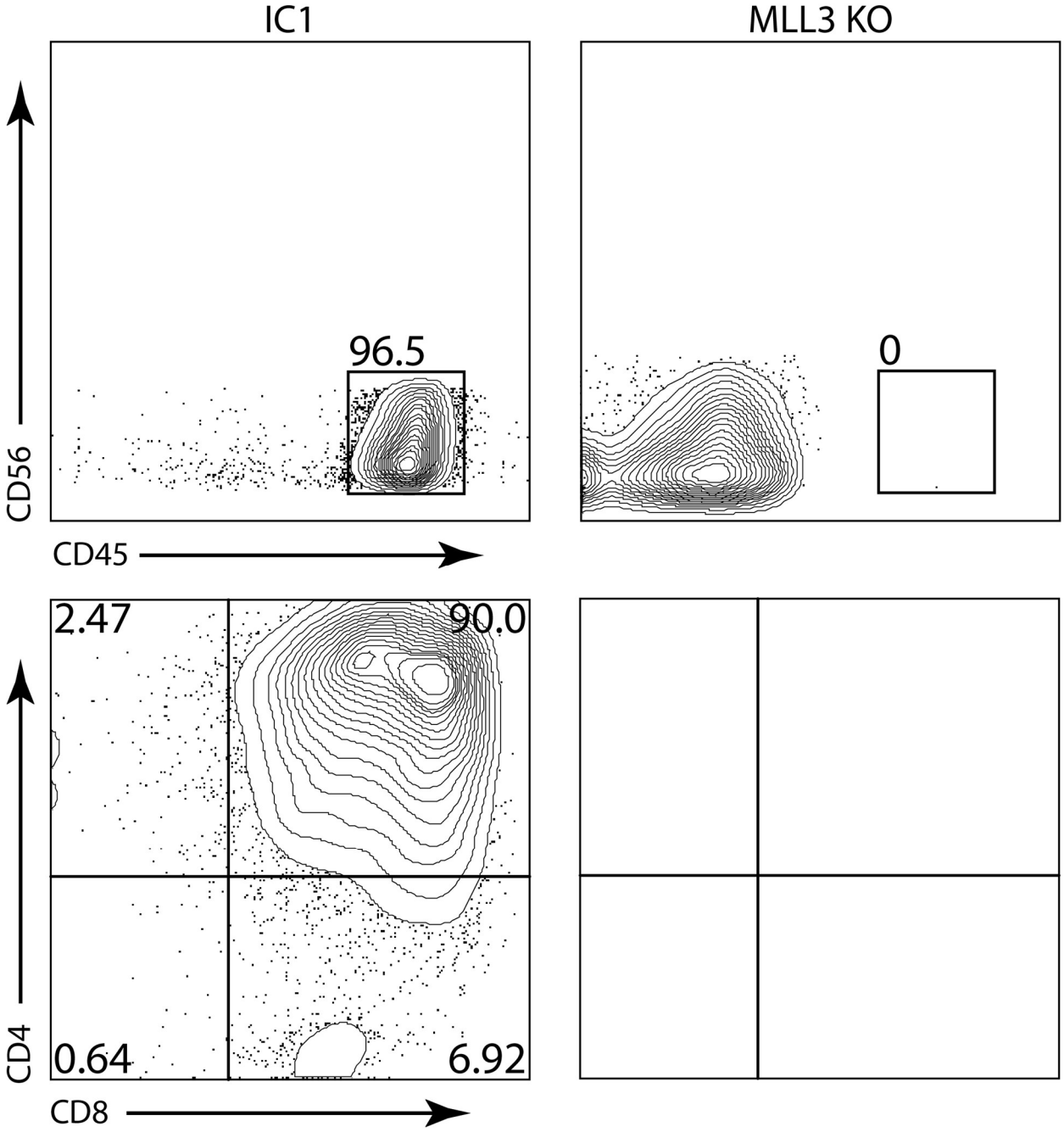


Figure 4.9: Day 3 flow cytometry in KMT2C KO cells. At day 3 the KMT2C cells displayed the expected phenotypes and were indistinguishable from the control line using this measure at this time point.

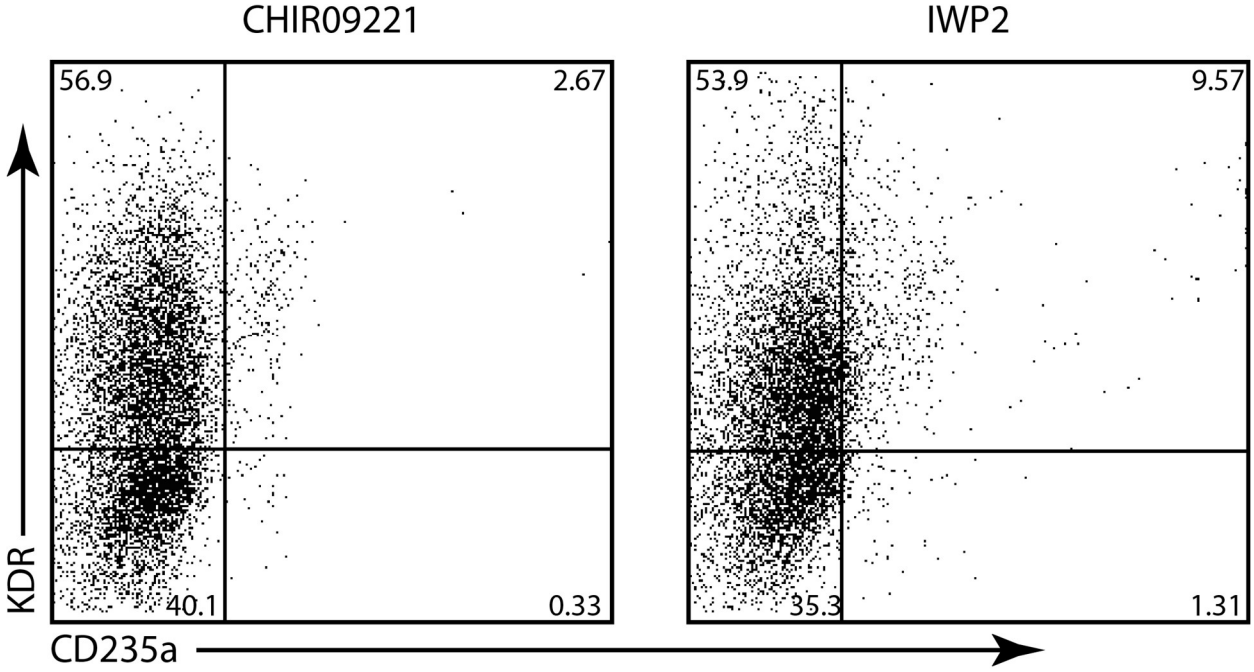
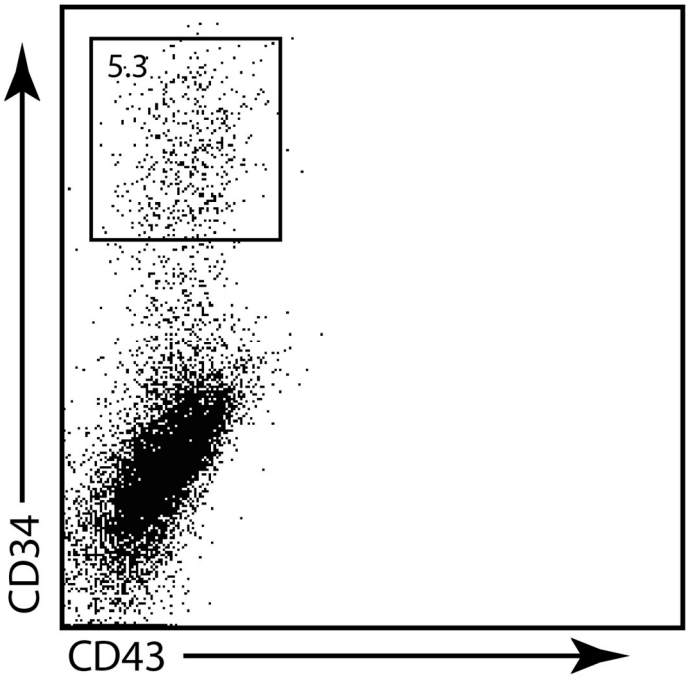


Figure 4.10: Day 8 flow cytometry in KMT2C KO cells. By flow cytometry, the day 8 KMT2C KO cultures displayed the expected cell surface markers. The populations were comparable in size to the same populations in control cells for a given assay.

a)



b)

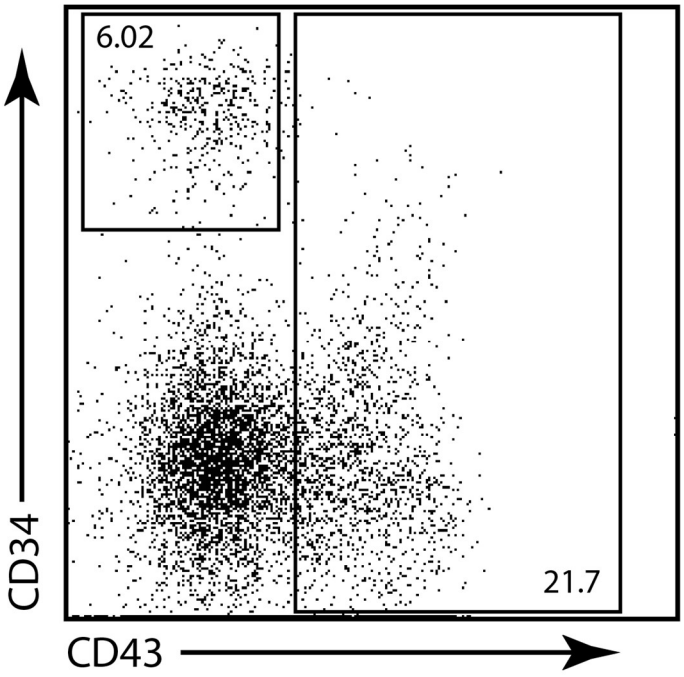


Figure 4.11: KMT2C (MLL3) KO cells fail to undergo EHT. Despite having a CD34+,CD43- endothelial population that was similar to that of control cells, the KMT2C KO cells did not undergo EHT. These cultures were watched for several more days to see if there was simply a timing difference between the two lines, but even after weeks in culture there was no indication of EHT in KMT2C KO cells.

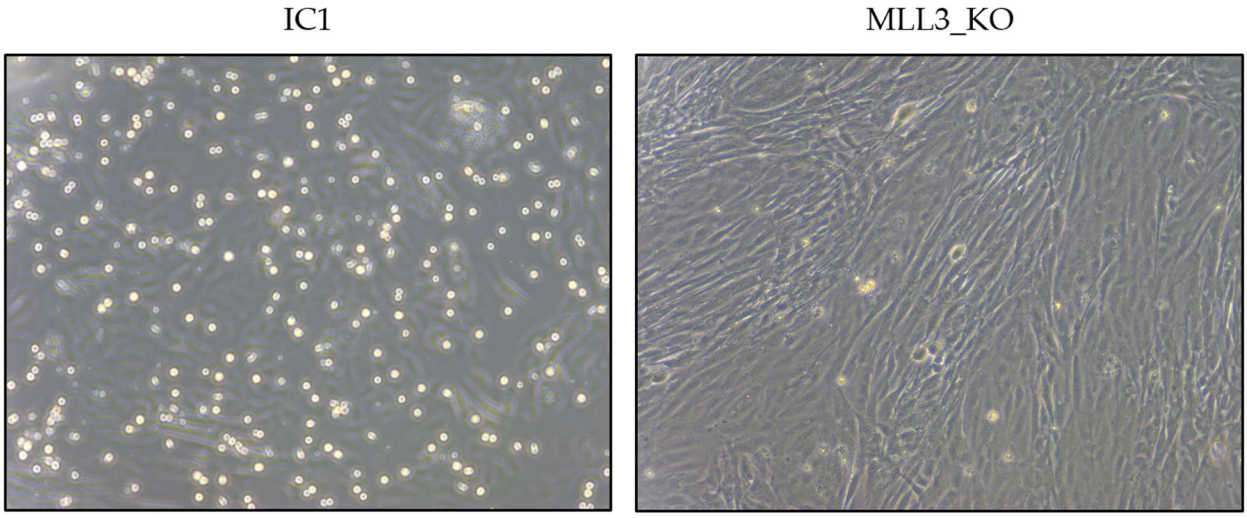


Figure 4.12: Day 8 flow cytometry in control and KMT2C KO cells. The addition of two markers allowing differentiation of various types of endothelium clearly demonstrates that the HE (red box) is not present in the KMT2C KO cells, despite similar amount of arterial endothelium (purple box) in both cultures.

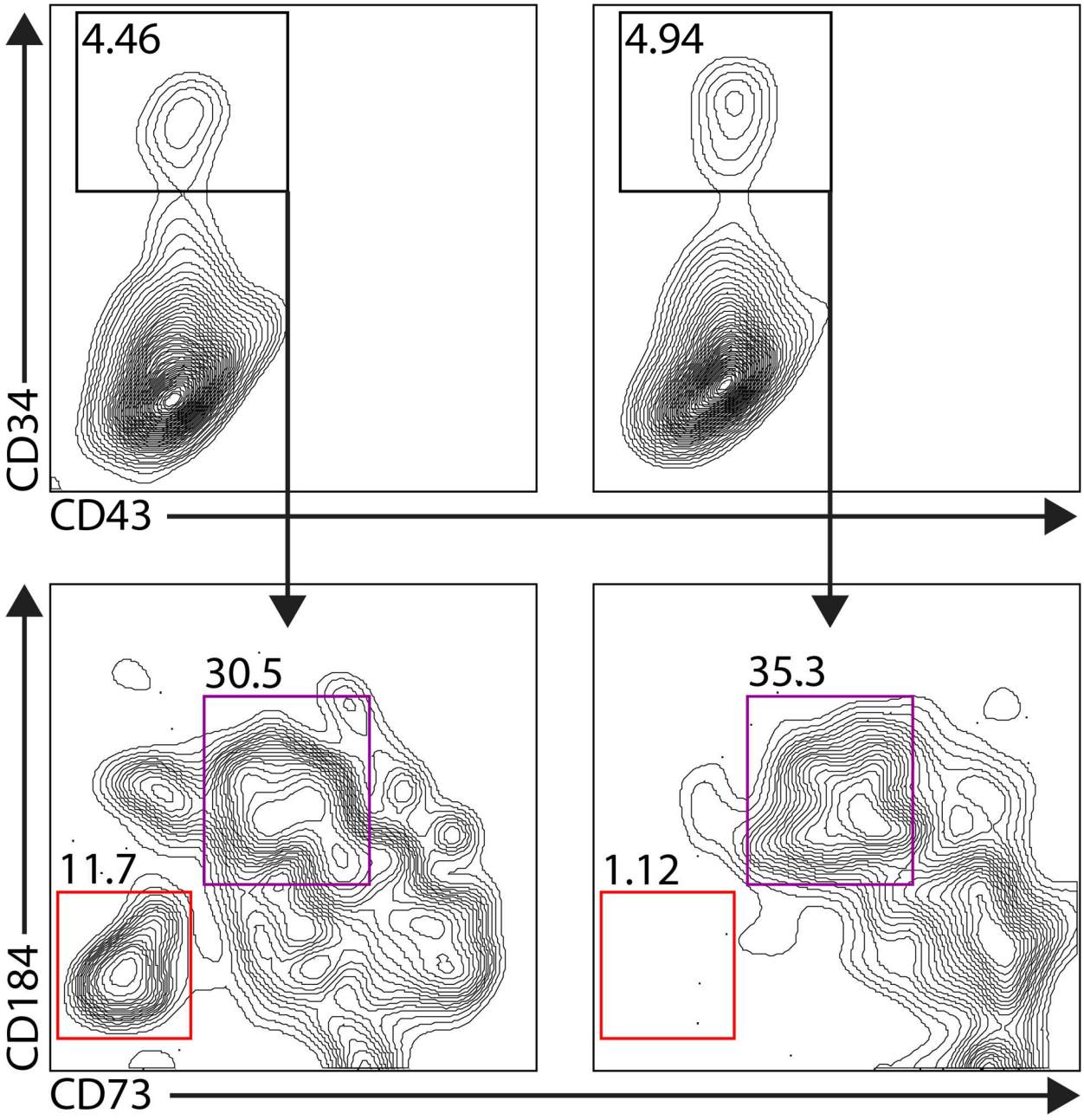


Figure 4.13: Definitive colony-forming assays. Definitive-specified KMT2C KO cells do not form colonies, indicating that the few cells that are present in the HE gates lack any hemogenic potential.

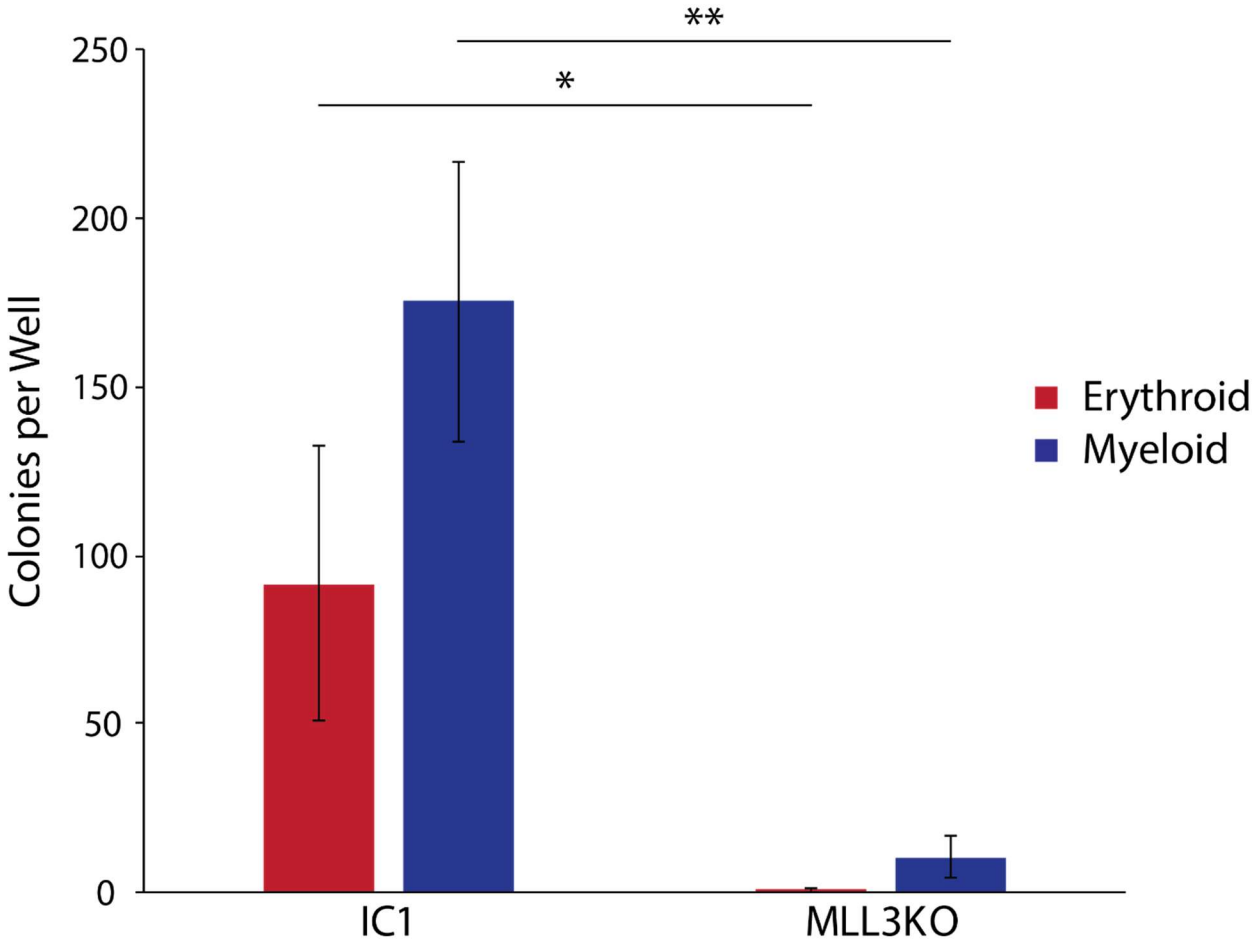


Table 4.1 Day 0 Medium

Component	Concentration
SFD	--
Glutamine	1%
Transferrin	150 ug/mL
Ascorbic Acid	50 ng/mL
MTG	11.25 ug/mL
BMP4	10 ng/mL

Table 4.2 Day 2 Medium

Component	Concentration
SFD	--
Glutamine	1%
Transferrin	150 ug/mL
Ascorbic Acid	50 ng/mL
MTG	11.25 ug/mL
BMP4	10 ng/mL
Activin A	0.3 ng/mL
bFGF	5 ng/mL
CHIR09921/IWP2	3 uM

Table 4.3 Day 3 Medium

Component	Concentration
StemPro34	--
Glutamine	1%
Transferrin	150 ug/mL
Ascorbic Acid	50 ng/mL
MTG	11.25 ug/mL
VEGF	15 ng/mL
bFGF	5 ng/mL

Table 4.4 Day 6 Medium

Component	Concentration
StemPro34	--
Glutamine	1%
Transferrin	150 ug/mL
Ascorbic Acid	50 ng/mL
MTG	11.25 ug/mL
VEGF	15 ng/mL
bFGF	5 ng/mL
IL6	10 ng/mL
SCF	50 ng/mL
IL11	5 ng/mL
EPO	2 U/mL
IGF1	25 ng/mL

Table 4.5 Day 8+ Medium

Component	Concentration
StemPro34	--
Glutamine	1%
Transferrin	150 ug/mL
Ascorbic Acid	50 ng/mL
MTG	11.25 ug/mL
TPO	30 ug/mL
IL3	30 ug/mL
SCF	100 ug/mL
EPO	4 U/mL
FLT3L	10 ug/mL
IGF1	25 ng/mL
IL6	5 ug/mL
IL11	5 ug/mL

Table 4.6 Serum Free Differentiation (SFD) Medium

Component	Amount
IMDM	750 mL
Ham's F12	250 mL
10% BSA – Fraction V	5 mL
B27 Supplement	10 mL
N2 Supplement	5 mL

Table 4.7 Antibodies

Target	Fluorphore	Conentration
CD34	APC	1:100
CD43	FITC	15:100
CD73	PE	2:100
CD184	BV421	2:100
CD235a	APC	1:100
KDR	Biotin (PE-SA)	15:100 (1:100)

5. Transcriptional and Epigenetic Profiling of hPSCs and Differentiating Hematopoietic Cells

5.1 Introduction

KMT2C deficient cells have a marked impairment in primitive and definitive hematopoietic differentiation. This differentiation process starts from an hPSC, continues through a mesodermal intermediate, an endothelial intermediate and culminates in hematopoietic progenitors and mature hematopoietic cells[84,93,144,153]. At each of these steps there is a cell-type specific transcriptional and epigenetic program. These programs determine the cell's identity, future potential and responsiveness to intrinsic and extrinsic stimuli. As the differentiation process progresses, various genes will need to be up- or down- regulated in a specific order and at specific times. This transcriptional control is effected by transcription factors and epigenetic modifiers, which will make genes more or less accessible by modifying chromatin accessibility, promoter and enhancer activation and possibly other factors.

KMT2C is, among other functions, a histone monomethylase with specificity for lysine 4 on histone 3. This mark (H3K4me1) is associated with active enhancers and areas of open chromatin. The enhancers regions targeted by KMT2C will have a positive impact on transcription and upregulate target genes. KMT2C has been incompletely characterized, however, and has several other domains that likely impart other functions to this very large

protein. Indeed, there is some evidence that KMT2C is capable of binding DNA directly, acting as a ubiquitin ligase, and responding to retinoic acid signaling in addition to methyltransferase activities. This wide range of potential activities explains the severe phenotype that we observed as well as the other phenotypes that have been associated with KMT2C dysfunction.

Recognizing that KMT2C would likely influence a number of transcriptional factors and epigenetic features during hematopoietic differentiation, we interrogated a number of these, including transcript abundance via RNA-seq, chromatin accessibility via ATAC-seq, and H3K4 monomethylation via ChIPmentation. We performed these at various timepoint throughout the differentiation process because the effect of KMT2C deletion might act at each independently or progressively as the process moves forward.

5.2 Methods

5.2.1 Cell sorting and Isolation

Cells were dissociated and stained as previously described (Section 4.2). Sorting was performed on a BD Aria FACS machine. For day 3 mesoderm samples, cells that were KDR+ and CD235a- were kept. For RNA-seq all of the cells that matched the phenotype of interest were kept, yields ranged from 3x10^5 – 1.5x10^6. For ATAC-seq and ChIPmentation, cells were split into 50,000 cell aliquots before processing.

5.2.2 RNA isolation

Sorted cells were washed and resuspended in 1 mL TRIzol reagent and frozen at -80C until processed. When all the samples had been collected, the -80C TRIzol samples were thawed at room temperature. 200 uL chloroform were added to each sample which was then mixed and

incubated at room temperature for 5 minutes. Samples were centrifuged for 15 minutes at 12000g, 4C. The upper aqueous phase was transferred to a fresh 1.5 mL tube, and 500 uL isopropanol was added. After a 10-minute incubation at room temperature, samples were centrifuged at 12000g for 10 minutes at 4C. The supernatant was discarded and the RNA pellet was washed once in 1 mL 70% ethanol followed by centrifugation at 7500g for 10 minutes at 4C. The supernatant is discarded and the pellet is allowed to dry for 5-10 minutes at room temperature. The RNA is then resuspended in ddH2O. RNA yield and quality was determined by nanodrop.

5.2.3 RNA-seq prep

Total RNA was processed by first depleting ribosomal RNA using the Ribo-Zero kit (Illumina Cat. MRZH116). This step removes the 28S, 18S, 5.8S, and 5S rRNA from solution, leaving a vastly enriched pool of mRNA. The mRNA is reverse transcribed into cDNA using the Sigma Seqplex kit (Sigma Cat. No SEQR-10RXN). Library preparation follows cDNA preparation and the samples are then sequenced on the Illumina HiSeq 3000 sequencer.

5.2.4 RNA-seq analysis

Raw reads were pseudo-aligned using kallisto[154] and then differential gene expression analysis and transcript quantification was performed using sleuth[155]. Other analysis was performed in R (Version 3.2.1, available at www.cran.r-project.org).

Reads were also aligned for using STAR[156,157]. The .sam output from STAR was processed (sorted, indexed, converted to .bam) using samtools. The .bam files were visualized in the IGV genome viewer[158].

5.2.5 ATAC-seq library preparation

Sorted cells were washed then incubated for 5 minutes in lysis buffer. The lysed nuclei were then centrifuged at 500g for 15 minutes at 4C. After discarding the supernatant, the cells were resuspended in Transposase Buffer with 2 uL Transposase enzyme. After 1 hour incubation at 37C, the reaction was placed on ice, and cleaned up with a Qiagen MinElute kit. Amplify the DNA and add the necessary adapters and barcode with nine cycles of PCR. Following PCR perform SPRI AMPure bead clean up to enrich for <= 600 bp products. Perform nine more PCR cycles to amplify the products further, and then follow with another AMPure clean up. The library can now be sequenced. A minimum of 30 million reads should be obtained for each sample.

5.2.6 ATAC-seq analysis

Raw reads were aligned to hg19 using bowtie2, option –X 2000. The resulting alignments were sorted, indexed, filtered and converted to .bam format using samtools. The filtering step removed reads mapping to blacklisted regions[159], the mitochondrial genome and sex chromosomes. Picard-tools (http://broadinstitute.github.io/picard) removeDuplicates was used to eliminate PCR duplicates. Homer was then used to create tag directories. Peaks were called using MACS 2[160] and annotated with Homer[161]. ATAC-seq peak visualization was performed with IGV and subsequent analysis was performed in the R software package.

5.2.7 ChIPmentation

A detailed version of this protocol is available[162]. Wash cells once in PBS and fix in 1% paraformaldehyde for 10 minutes at room temperature. Stop the reaction by adding glycine. Centrifuge 5 minutes at 500 x g at 4C to collect the cells. Wash twice with ice-cold PBS, then

resuspend in RIPA buffer and incubate 10 minutes on ice. To isolate the nuclei, centrifuge 10 minutes at 1000 x g, 4C. Samples were sonicated on a Covaris S220 to obtain fragment 200-700 bp in length. Spin the samples at full speed for 5 minutes, 4C. Increase the volume to 200 uL and incubate with H3K4me1 antibody (Diagenode pAb-196-050) overnight at 4C. Concurrently, block Protein A coated magnetic beads overnight in in 0.1% BSA in PBS. Add blocked beads to the IP fragments and incubate 2 hours at 4C. Wash beads in RIPA twice, RIPA-500 once and TE twice. Wash beads twice with cold Tris-HCl, then resuspend in 30 uL tagmentation reaction mix containing 1 uL Tagment DNA enzyme. Incubate 10 minutes at 37C. Wash twice with RIPA buffer. Reverse crosslinking by incubating in 70 uL Elution Buffer for 1 hour at 55C then 8 hours at 65C. Transfer supernatant to a new tube. Purify the DNA with SPRI AMPure beads (1:2 ratio).

5.2.8 ChIPmentation analysis

Raw reads were aligned to hg19 using bowtie2, option –X 2000. The resulting alignments were sorted, indexed, filtered and converted to .bam format using samtools. The filtering step removed reads mapping to blacklisted regions[159], the mitochondrial genome and sex chromosomes. Picard-tools (http://broadinstitute.github.io/picard) removeDuplicates was used to eliminate PCR duplicates. Homer[161] was then used to create tag directories. Peaks were called using MACS 2[160] and annotated with Homer. ChIPmentation peak visualization was performed with IGV and subsequent analysis was performed in the R software package.

5.3 Results

KMT2C KO cells have the most pronounced phenotype on day 8 in definitive-specified cultures, so we started our analysis here. Since KMT2C KO cells do not generate any HE cells, we chose to characterize the CD34+, CD43- mixed endothelial population. We sorted this population (Figure 3.1) and performed RNA-seq, ATAC-seq and ChIPmentation for H3K4me1. Our initial analysis focused on the RNA-seq data. Differential expression analysis provided a number of transcripts that were significantly differentially expressed. We plotted the p-value from this analysis against the fold change in a given transcript to yield the volcano plot in Figure 5.2. Although we knew that KMT2C was a positive regulator of transcription we were surprised to find that the vast majority of transcripts that were upregulated and significantly differentially expressed were higher in the control cells. By contrast very few transcripts were upregulated in the KMT2C cells were upregulated. The few genes that were upregulated in KMT2C cells tended to be upregulated to a much lesser extent. This is consistent with the role of KMT2C, and suggests that the hematopoietic defects in KMT2C KO cells is a result of their inability to establish the relevant transcriptional profile for hematopoietic maturation.

We further explored this data set by selecting the transcripts that exceeded a fold-change and significance based filter (Figure 5.2). The genes that met these criteria for either KMT2C KO cells or control cells were used as inputs into gene set enrichment software to determine the pathways or processes that were overrepresented in these data (Tables 5.1 and 5.2). As expected, the control cells were enriched in relevant pathways including Wnt signaling, notch signaling, specific HOX genes and BMP signaling (Figures 5.3,5.5-5.13). By contrast KMT2C KO cells

had upregulated transcripts important for neural development, heart development and SHH signaling (Figures 5.4-5.13). The latter is interesting as SHH signaling is important for a number of developmental processes including hematopoiesis. Presumably however, the timing or amount of expression of these genes was inappropriate for normal hematopoietic development.

The transcriptional identity of a cell is the result of several factors, including chromatin accessibility and enhancer activity. We next analyzed our ChIPmentation and ATAC-seq data and found that they strongly supported our RNA-seq findings. The KMT2C competent cells had the expected tens of thousands of peaks at H3K4me1 sites. In contrast to this, KMT2C KO cells had only a few thousand. Further, the read density around the control defined H3K4me1 peaks was much lower in the KMT2C KO cells (Figure 5.14). KMT2C is only one of several enzymes capable of placing H3K4me1 marks, so the peaks that remain in this dataset are presumably the result of other histone methyltrasnferases.

We extended these observations by intersecting the ATAC-seq defined peaks with the H3K4me1 peaks. To do so, we extended the region around the H3K4me1 peaks by 500 bp on either side. In this expanded window, we looked for the presence of ATAC-seq peaks. As expected, there a great deal of overlap between the H3K4me1 enhancer mark, and open chromatin (Table 5.3).

The difference in definitive-specified control and KMT2C KO cells is pronounced at the day 8 endothelial stage. We found that there is strong transcriptional and epigenetic support for this. However, it is entirely possible that the hematopoietic differentiation process was perturbed well before this stage. To explore this possibility, we sorted day 3 KDR+, CD235a- definitive specified mesoderm cells from both control and KMT2C KO cultures (Figure 5.16), and performed RNA-seq. As before, we generated a volcano plot using fold-change and the p-value

of a test of differential expression (Figure 5.17). We applied a similar fold-change and p-value based filter to generate a list of significant transcripts for downstream analysis. As before, there were more upregulated transcripts in the control cells than in the KMT2C knockout cells (Tables 5.4 and 5.5). However, the difference at day 3 was much smaller than the difference at day 8. This suggests that the deficit in KMT2C takes time exert an effect. It also potentially explains why KMT2C KO cells are relatively similar to control cells at this time point based on flow cytometry. Still, despite the similarity in KDR and CD235a expression, an enrichment analysis shows that the transcriptional profiles of these two cell lines are quite different even at this time point. The control cells expressed a collection of genes responsible for expected processes including axis specification, cardiovascular differentiation, gastrulation and organismal patterning (Figures 5.18,5.20-5.27). In contrast, the KMT2C KO mesoderm expressed a seemingly disordered collection of transcripts involved in neural and heart development (Figs 5.19-5.27). Further exploration of these observations is warranted, but as they stand, they are entirely consistent with our hypothesis of disordered gene expression and a failure to upregulate the relevant and necessary gene programs required for normal hematopoietic development.

5.4 Discussion

The two cell lines used in these studies differ only in their expression of KMT2C. Thus any epigenetic biases that are present between various hPSC lines are not relevant in this context. Thus, the marked difference in the ability of KMT2C competent and KMT2C deficient cells to undergo hematopoietic differentiation is entirely the result of KMT2C dysfunction. In this study we observed that KMT2C KO cells do not upregulate a large number of genes that are upregulated in control cells. These genes are enriched for important hematopoietic

differentiation pathways. The KMT2C KO cells do upregulate a smaller set of genetic programs that are unrelated to hematopoietic differentiation. This is consistent with the idea that these cells are unable to upregulate specific programs and instead undergo somewhat random differentiation that is not tethered to the developmental signals that are present. This effect was most pronounced at day 8, as was the phenotypic difference between control and knockout cells. However, there was a clear difference in the two cell lines as early as day 3 despite phenotypic similarity in flow cytometric assays.

ChIPmentation and ATAC-seq to explore activating histone marks and chromatin accessibility corroborated the transcriptional profiles observed in day 8 cells. KMT2C deficient cells lacked H3K4me1 enhancer marks and had much more closed chromatin relative to KMT2C expressing cells. Thus the differences in transcription were a result of epigenetic dysregulation on a genomic scale. Given the magnitude of these differences it is unsurprising that there was such a severe phenotype in differentiation assays. It is unknown what the specific effect of this dysregulation would be *in vivo,* but it is not difficult to bridge the gap between severe transcriptional and epigenetic dysregulation and the development of cancer. This is especially true when the dysregulation is the result of a single gene, and the IL patients have additional deleterious variants in other genes.

Figure 5.1: Day 8 Sorting Schema. The populations that were sorted for downstream analysis.

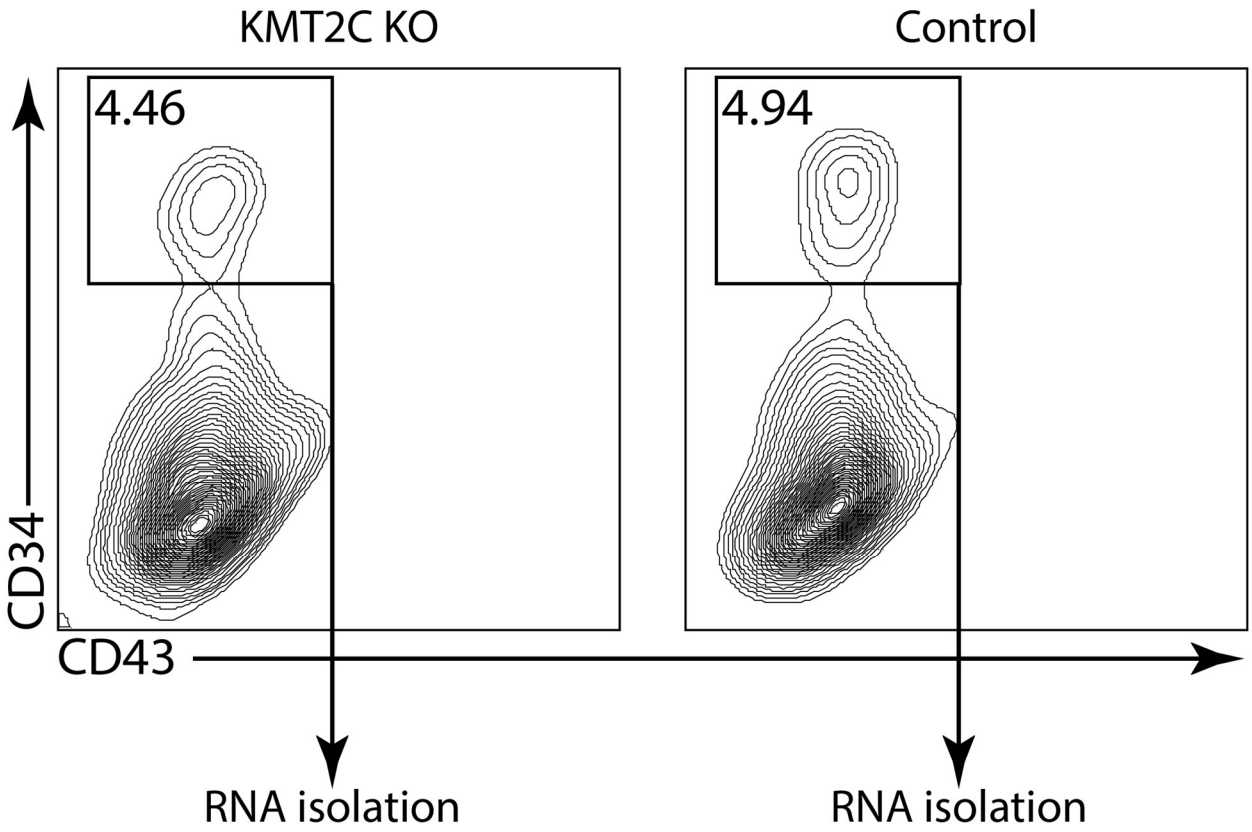


Figure 5.2 : Volcano plot of day 8 RNA-seq data. Each transcript was plotted by fold-change and p-value. The transcripts that exceeded a filter combining these two criteria were used as inputs for gene set enrichment analysis (red and green points)

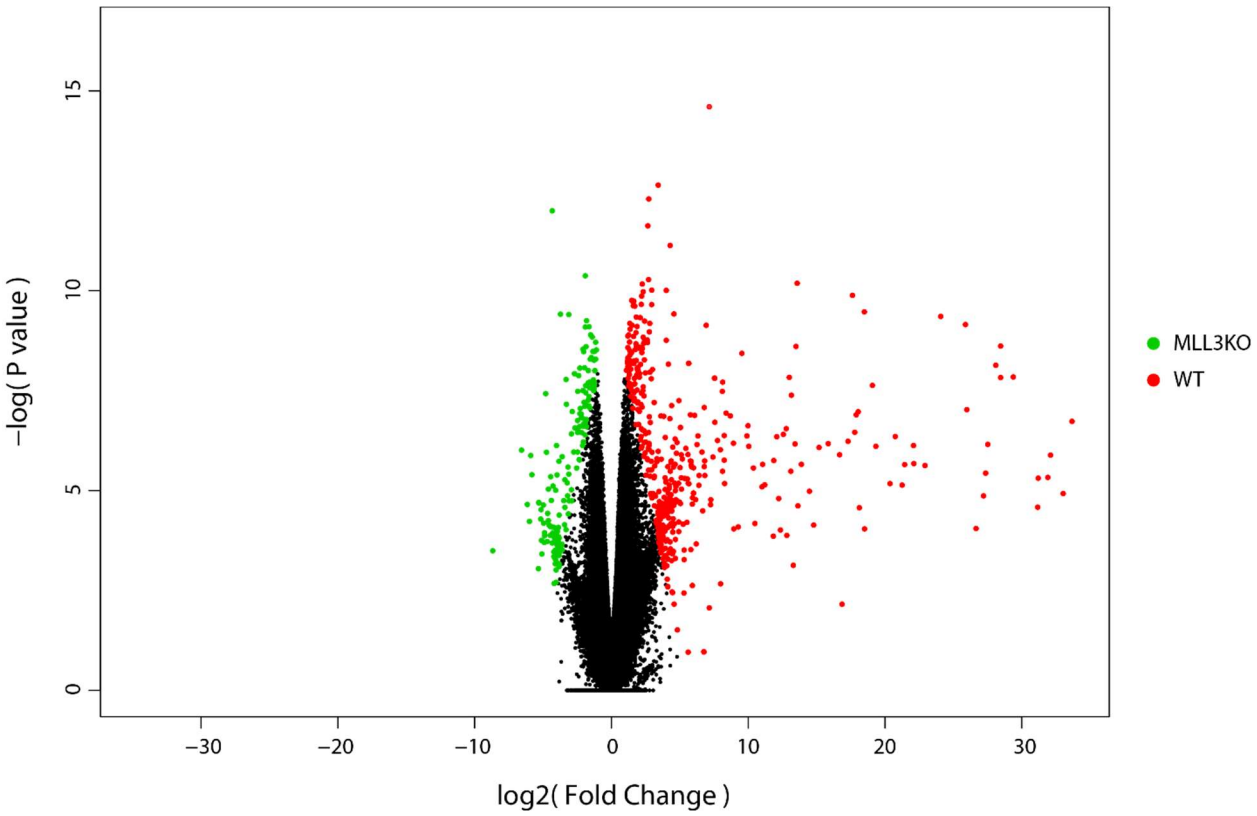


Figure 5.3 : Heatmap of transcripts upregulated in day 8 KMT2C KO cells. The transcripts included are listed in Table 5.1

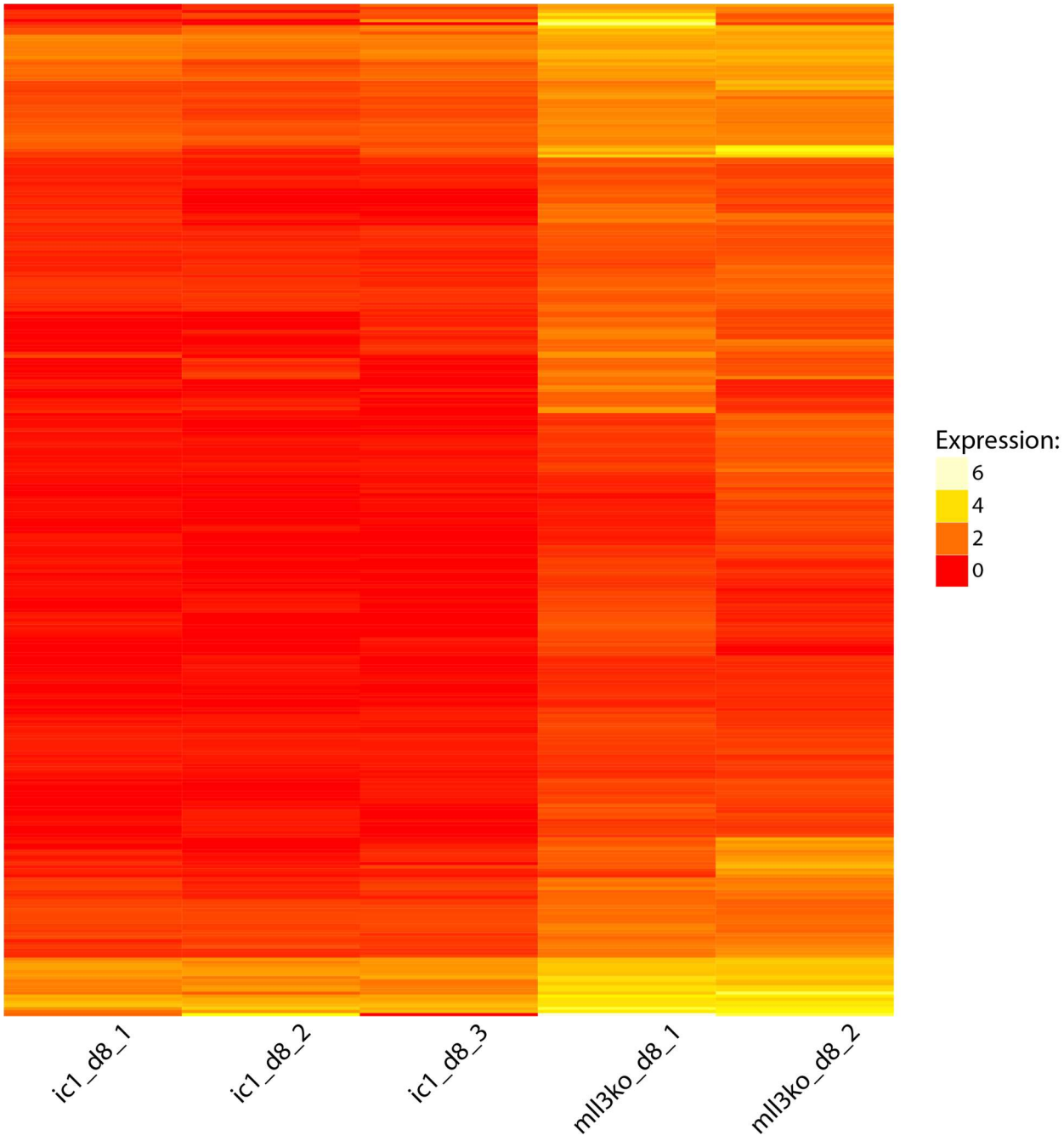


Figure 5.4 : Heatmap of transcripts upregulated in day 8 control cells. The transcripts included are listed in Table 5.2

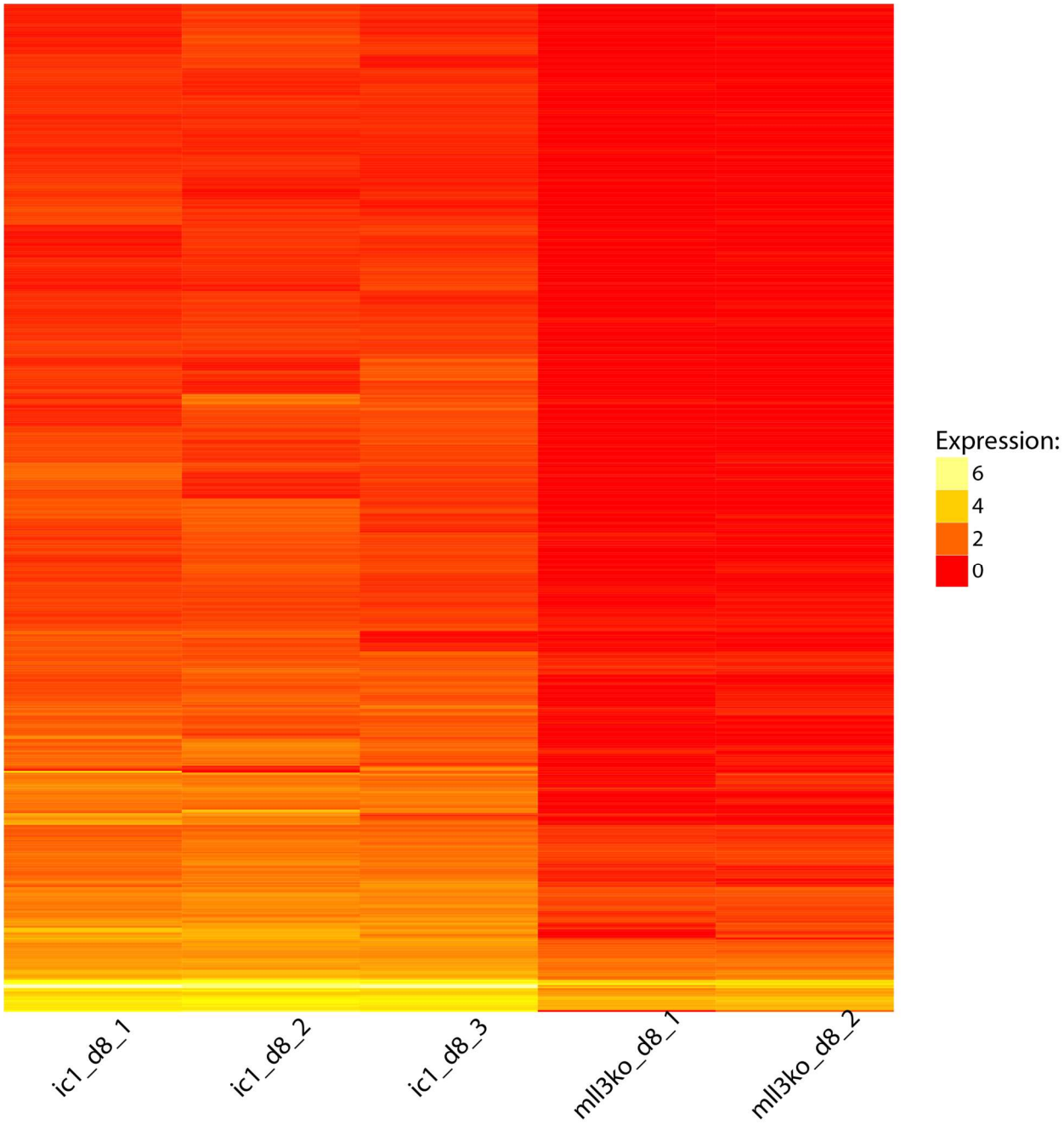


Figure 5.5: Heatmap of day 8 angiogenesis genes

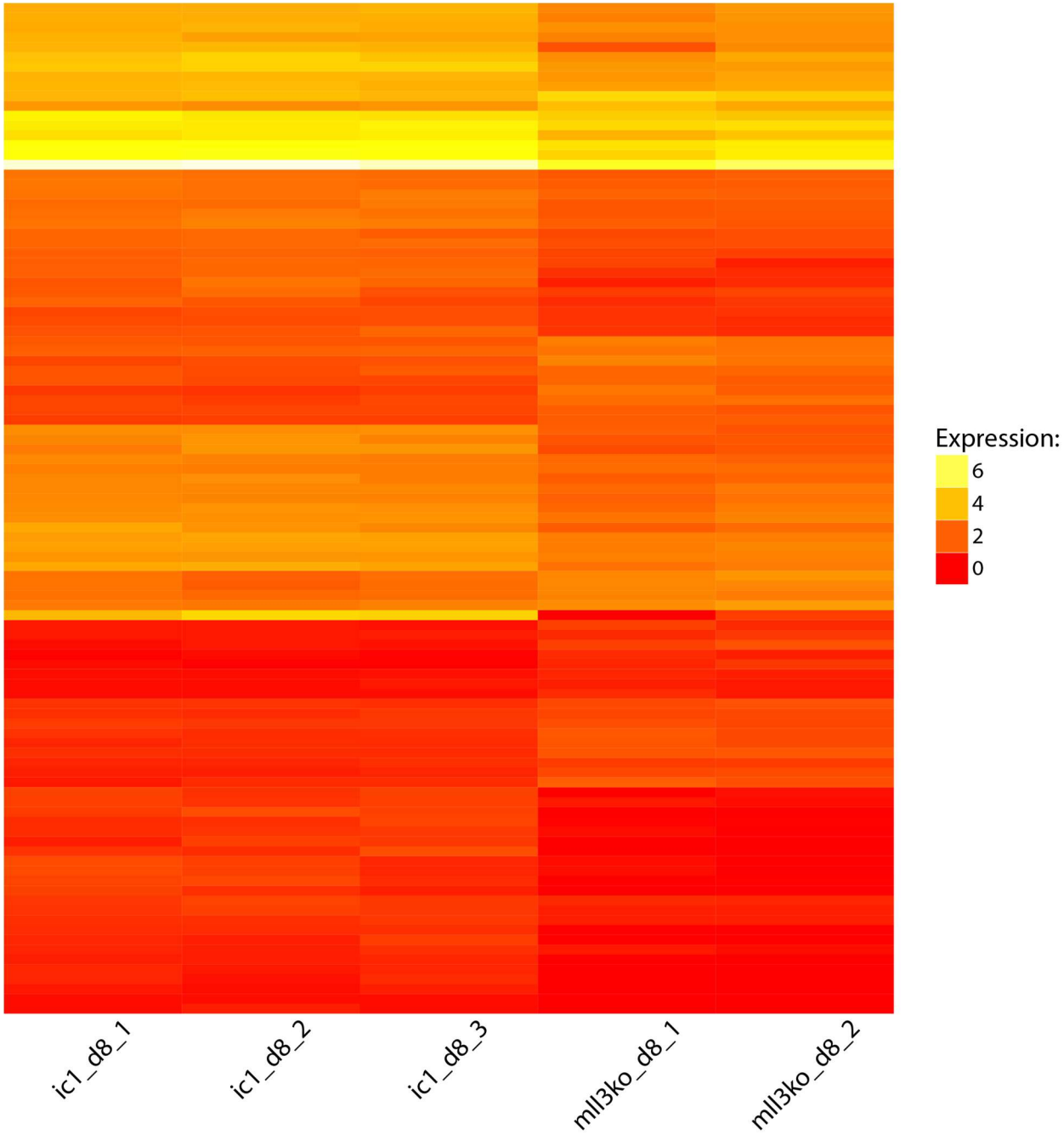


Figure 5.6: Heatmap of day 8 BMP genes

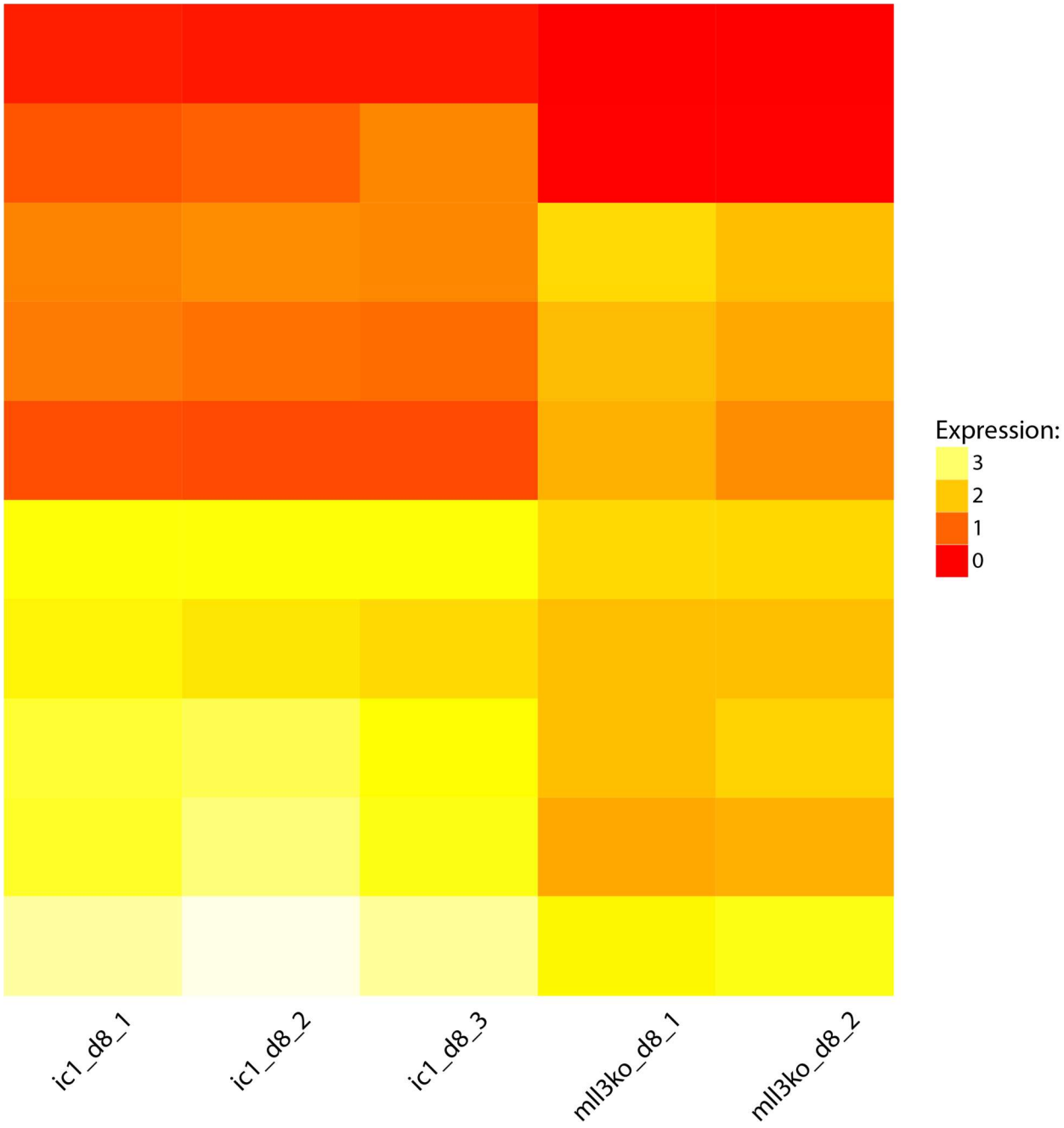


Figure 5.7: Heatmap of day 8 HOX genes

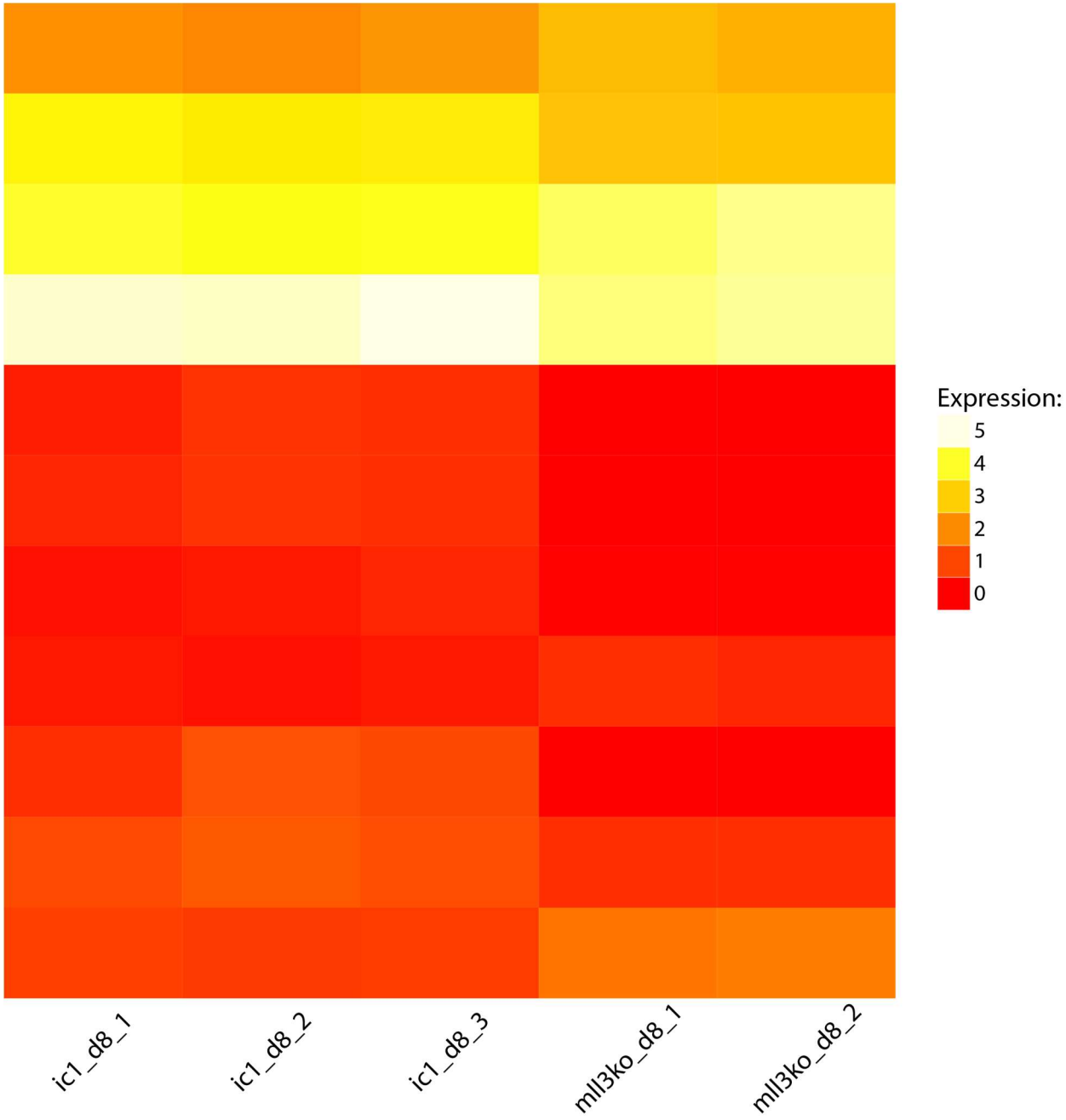


Figure 5.8: Heatmap of day 8 heart genes

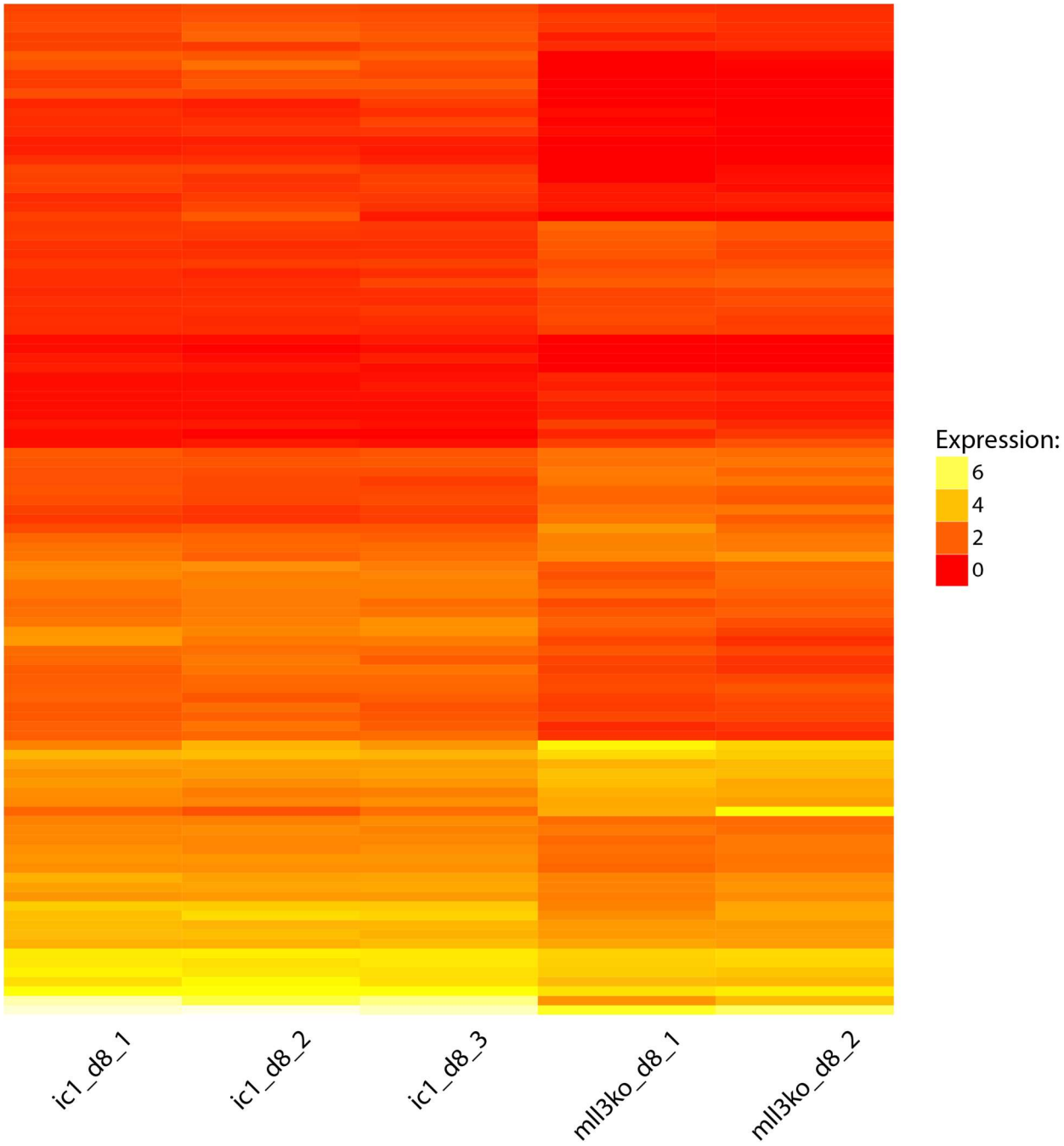


Figure 5.9: Heatmap of day 8 hedgehog genes

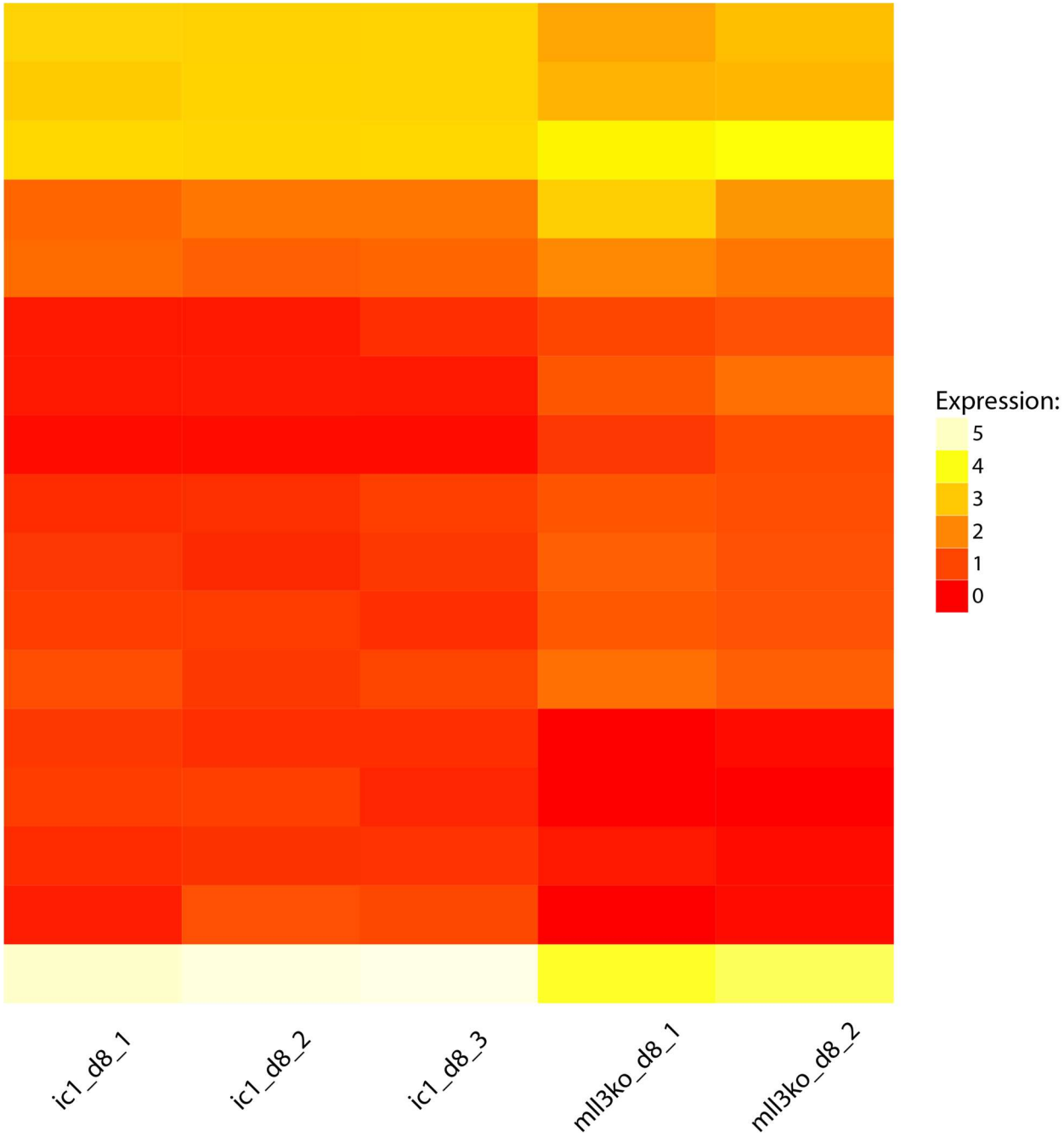


Figure 5.10: Heatmap of day 8 MAPK genes

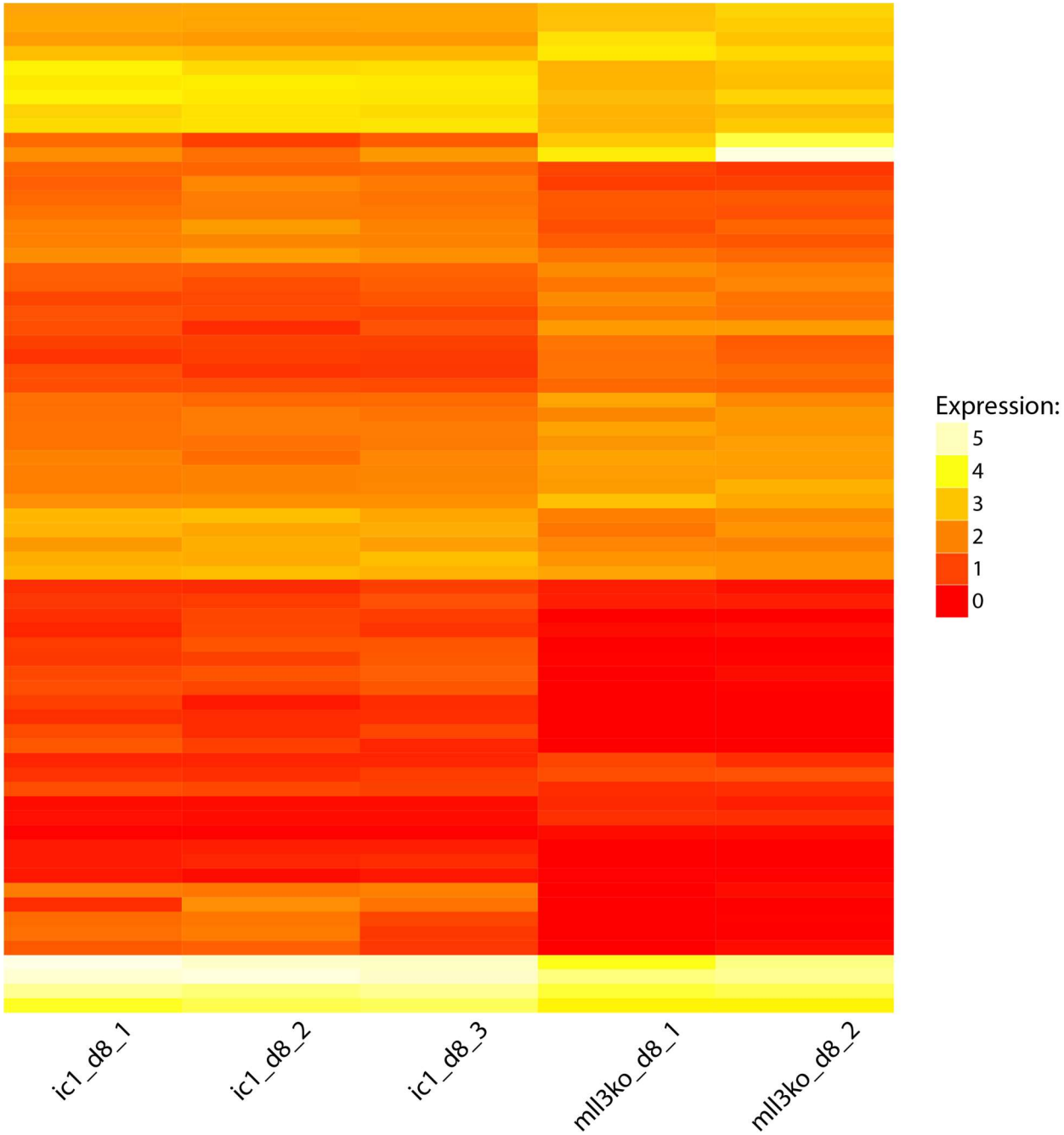


Figure 5.11: Heatmap of day 8 neuro genes

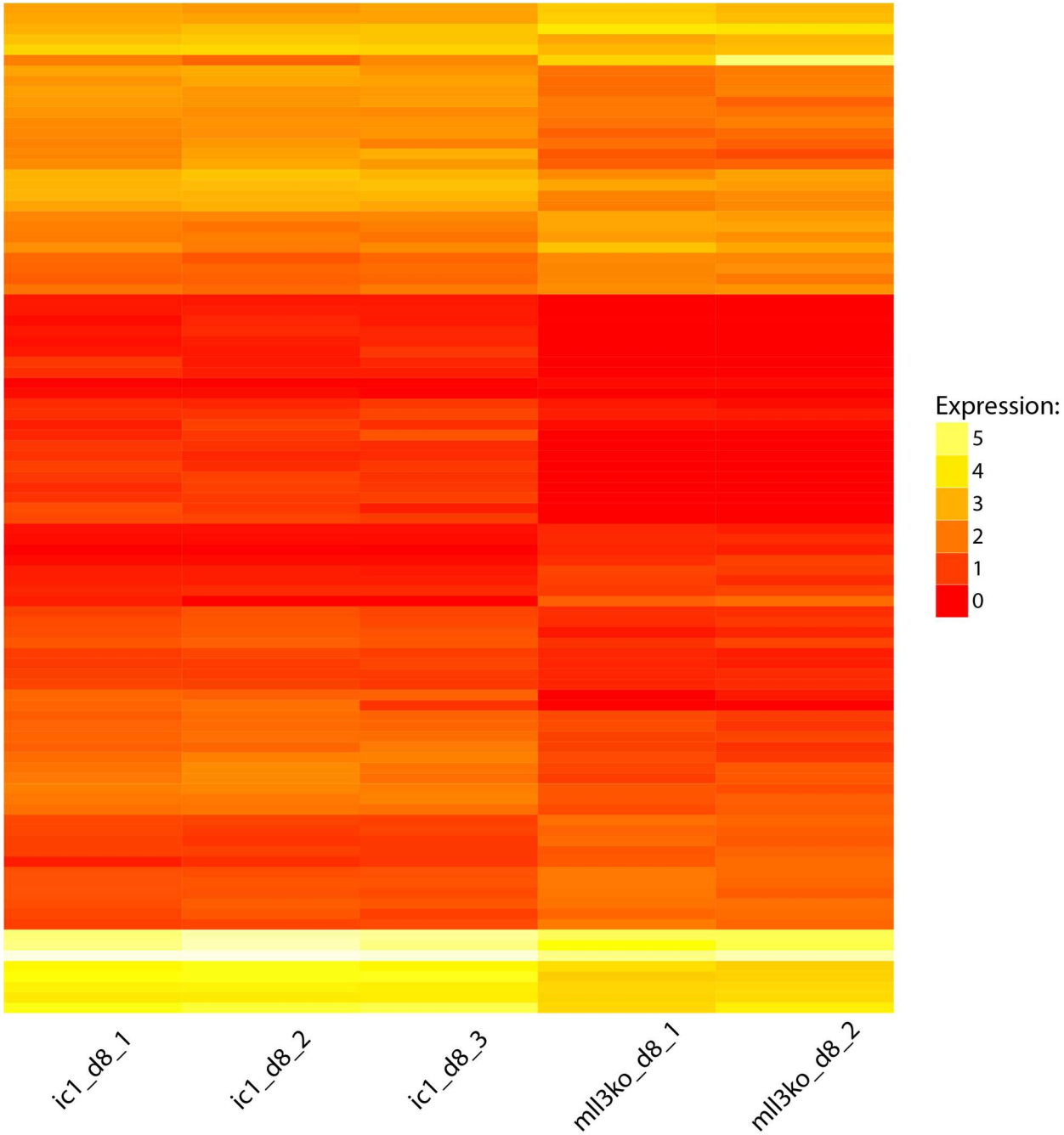


Figure 5.12: Heatmap of day 8 notch genes

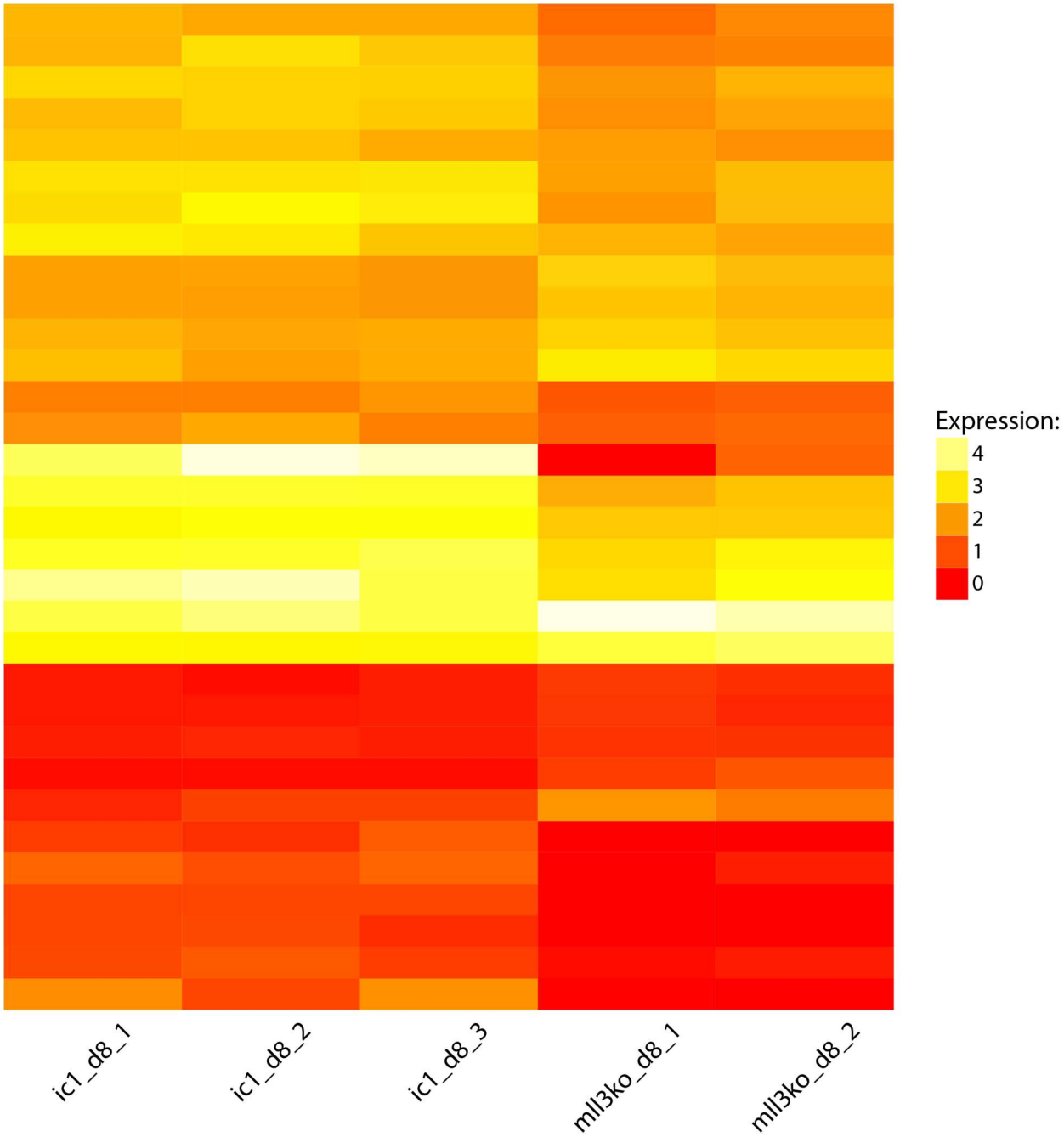


Figure 5.13: Heatmap of day 8 Wnt genes

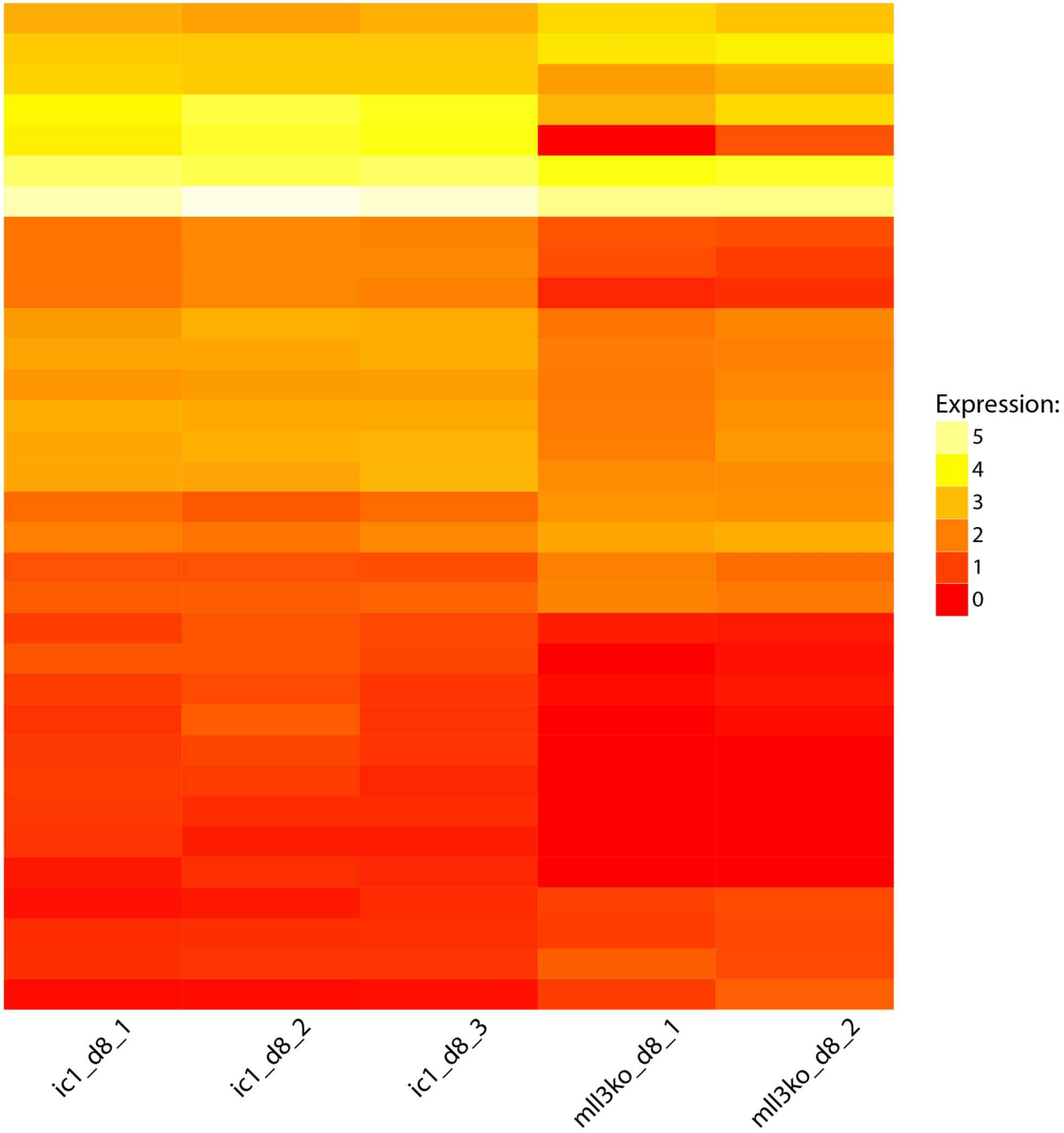


Figure 5.14: ChIPmentation peak heatmaps. The number of reads flanking all peaks called in control samples (columns 1 and 2) and KMT2C KO samples (columns 3 and 4) is depicted in the intensity of blue. There is a clear decrease in reads surrounding peaks in the KMT2C samples even in the peaks that are maintained, and a complete loss of most peaks.

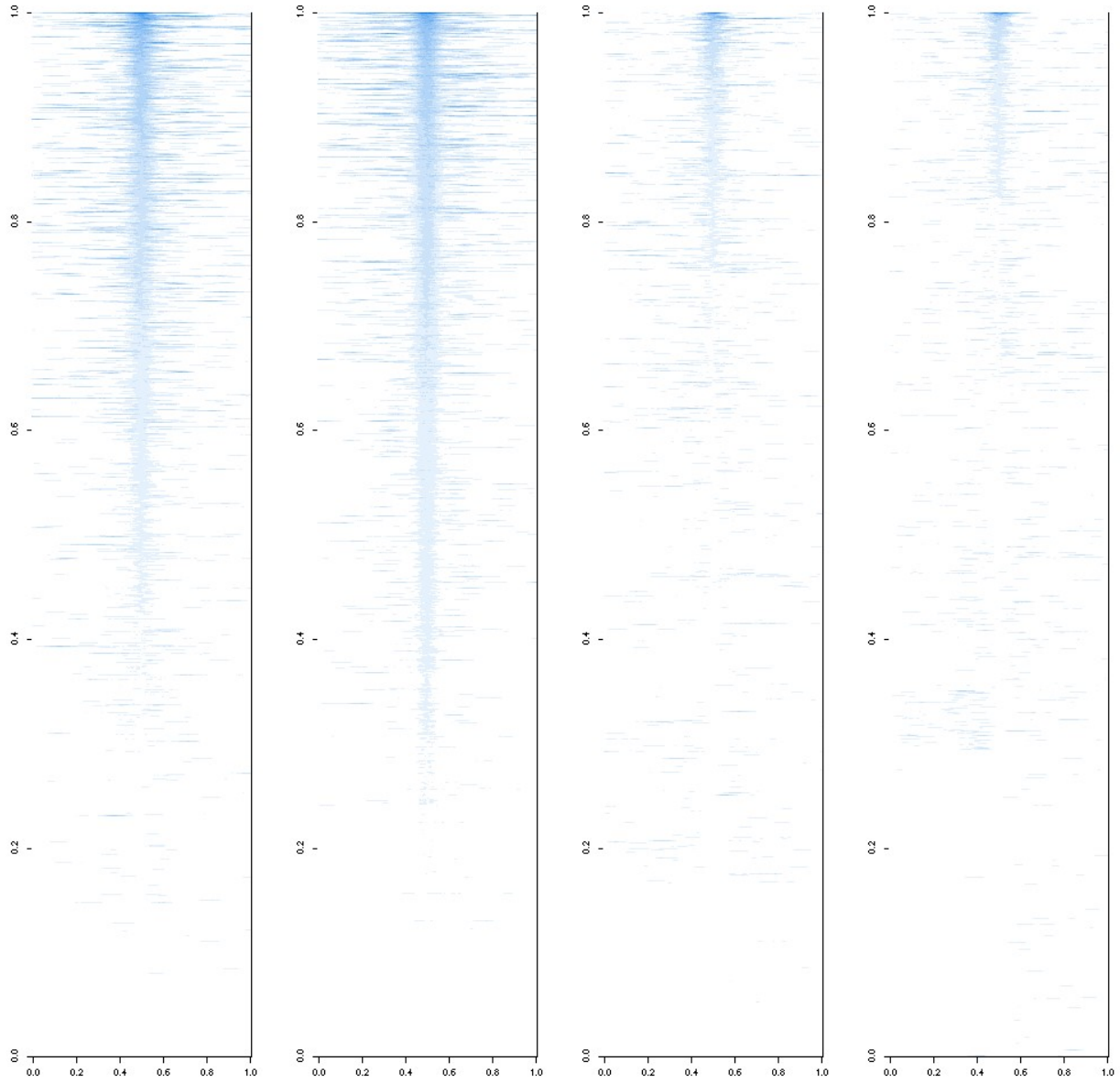


Figure 5.15: Sorting strategy for Day 3 RNA-seq.

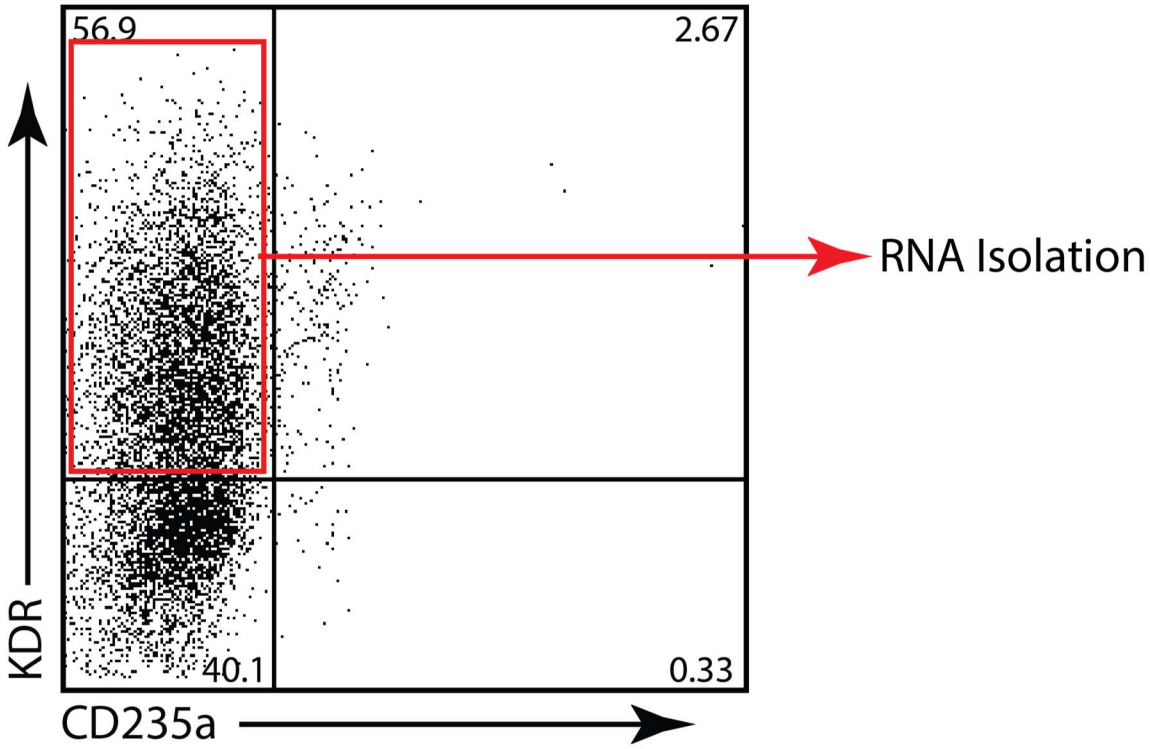


Figure 5.16: Volcano plot of d3 RNA-seq in KMT2C KO and control cells. The same plot as Figure 5.2, but for Day 3 RNA-seq data.

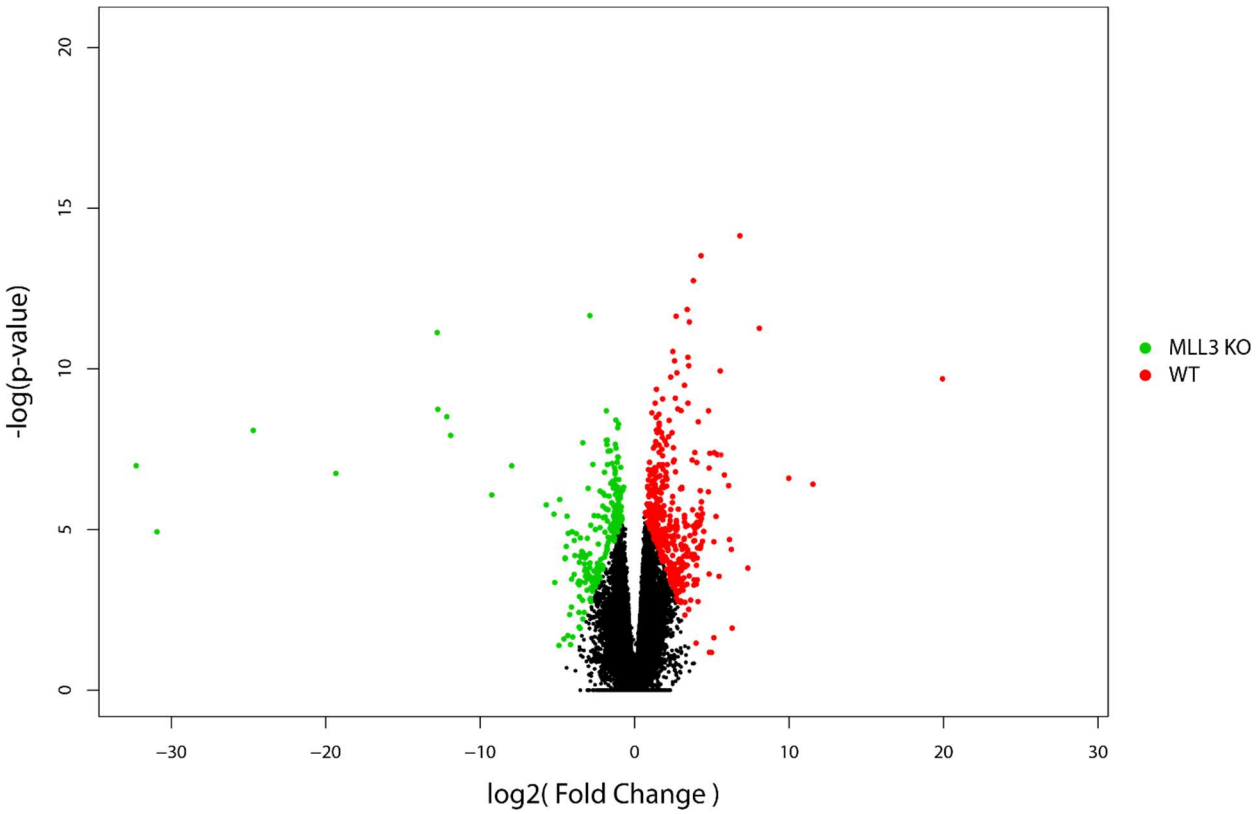


Figure 5.17: Heatmap of day 3 WT genes

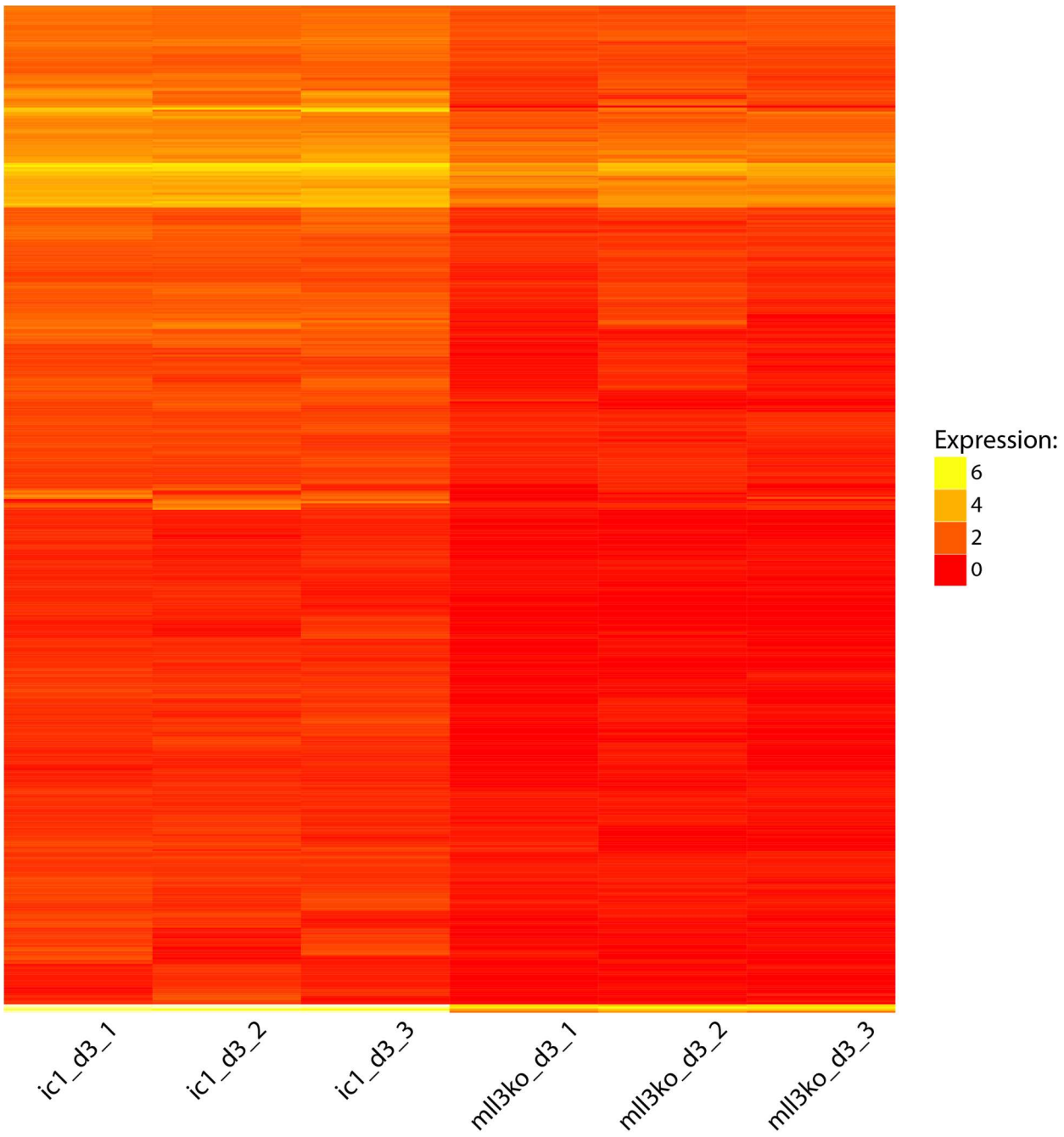


Figure 5.18: Heatmap of day 3 KO genes

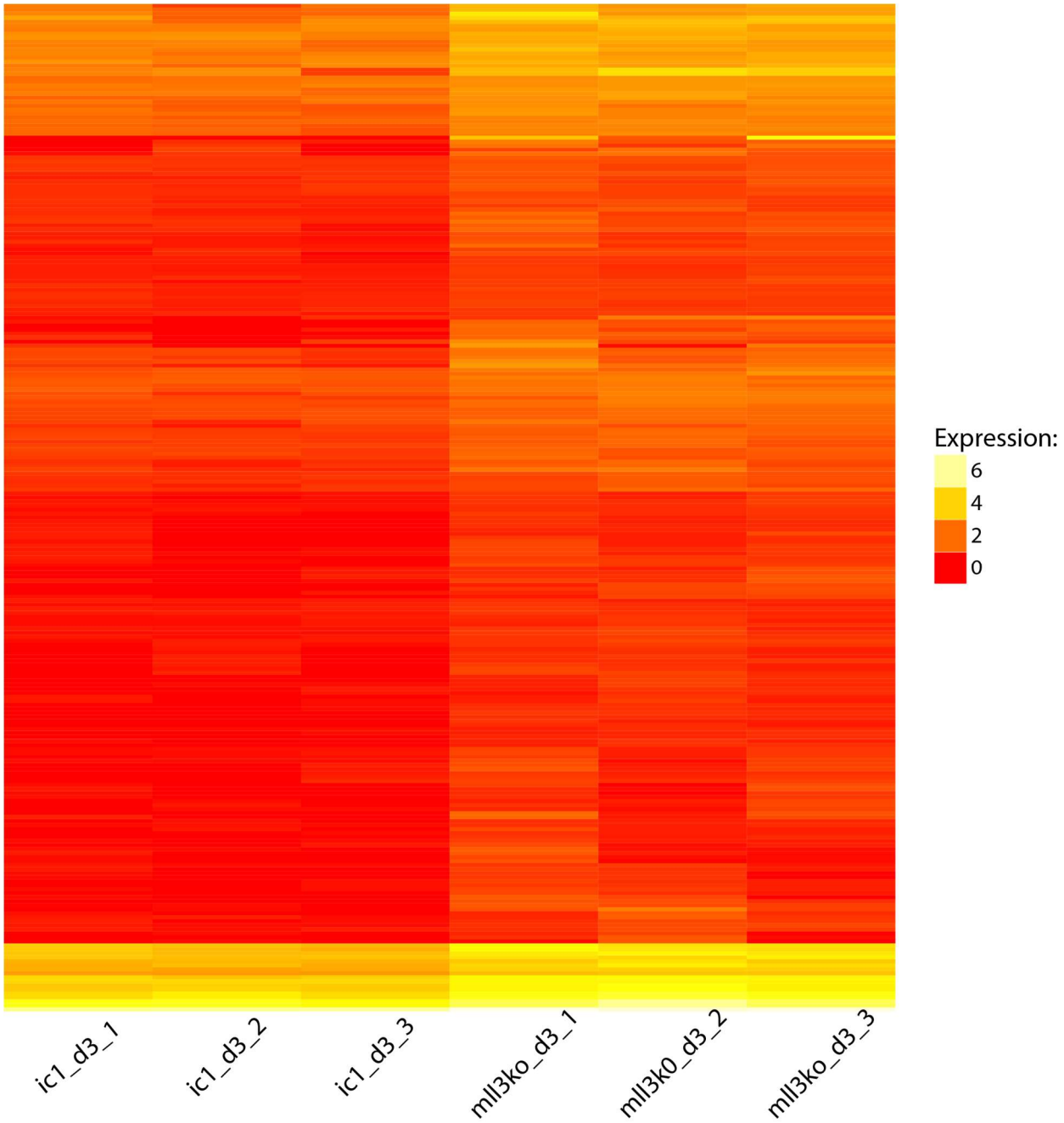


Figure 5.19: Heatmap of day 3 Axis Specification genes

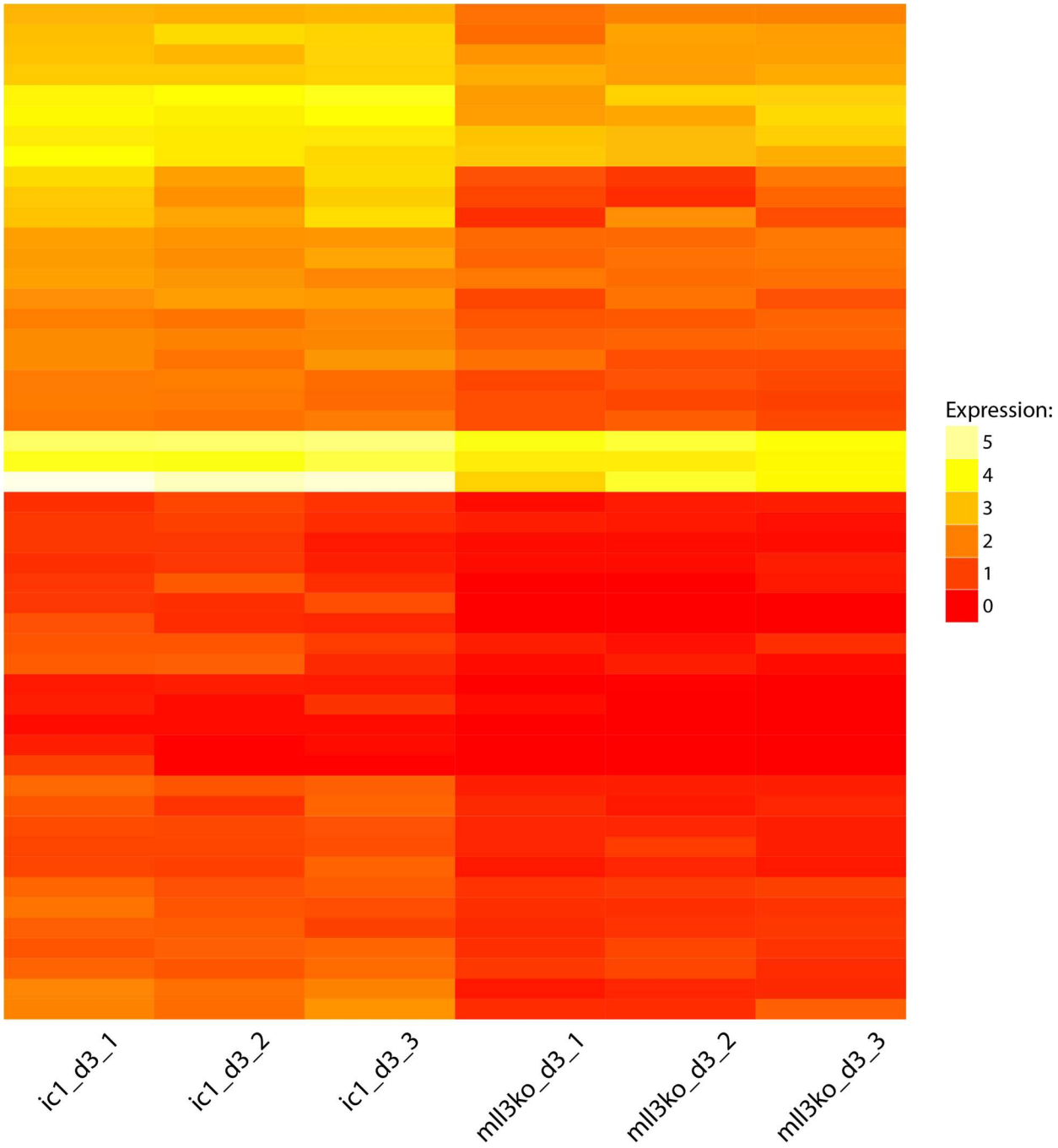


Figure 5.20: Heatmap of day 3 Cardiovascular Development genes

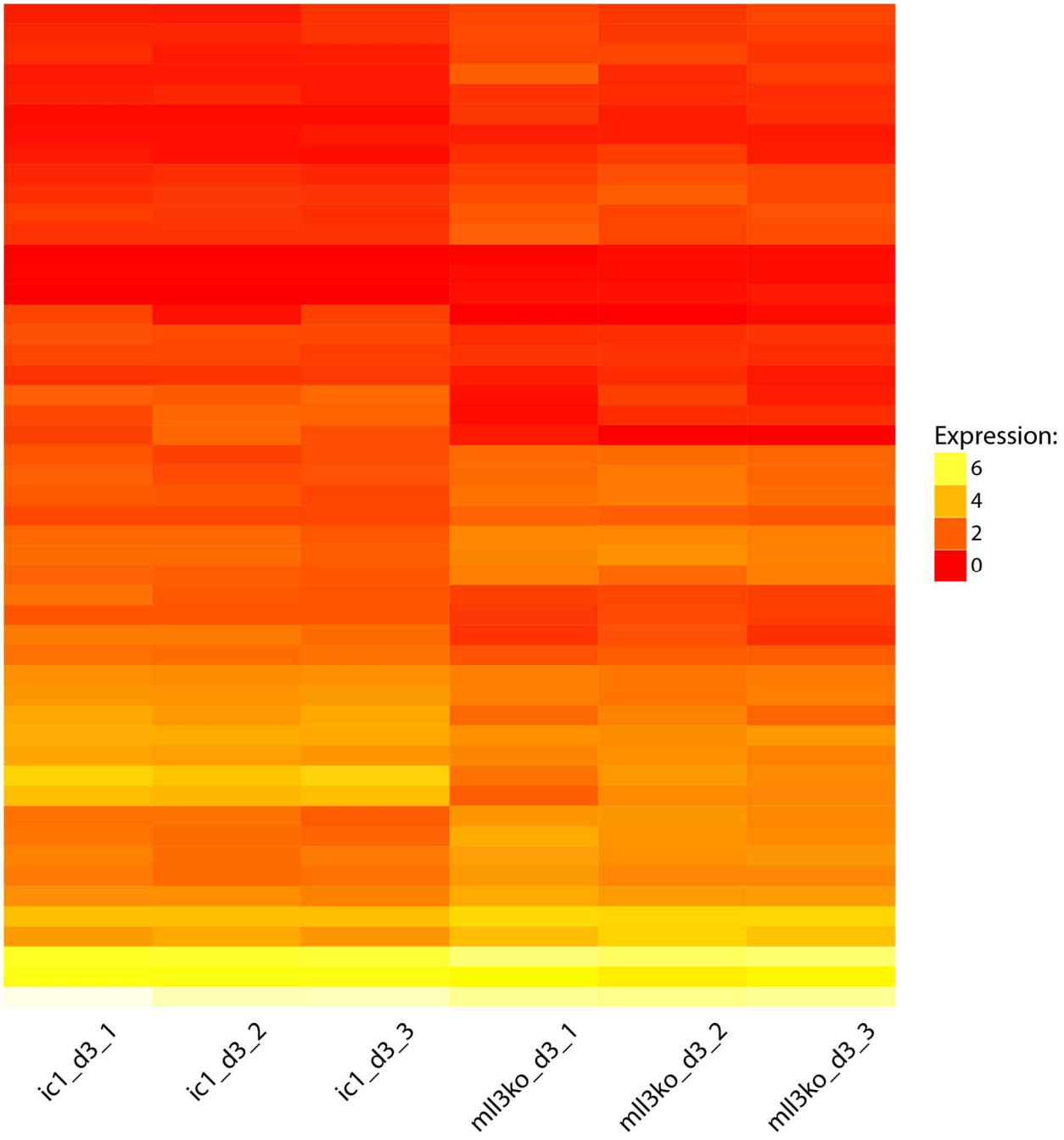


Figure 5.21: Heatmap of day 3 Organismal Differentiation genes

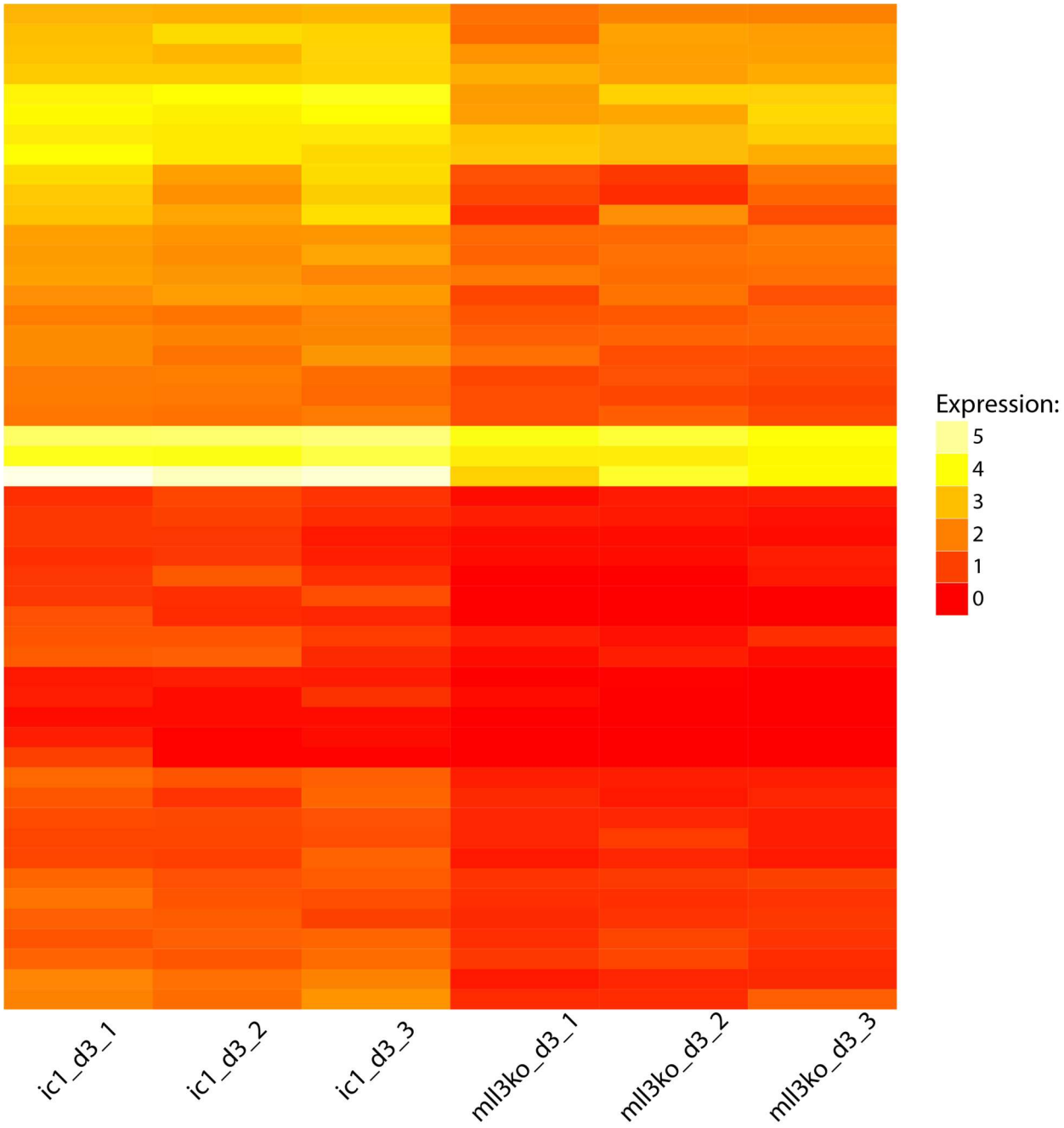


Figure 5.22: Heatmap of day 3 Gastrulation genes

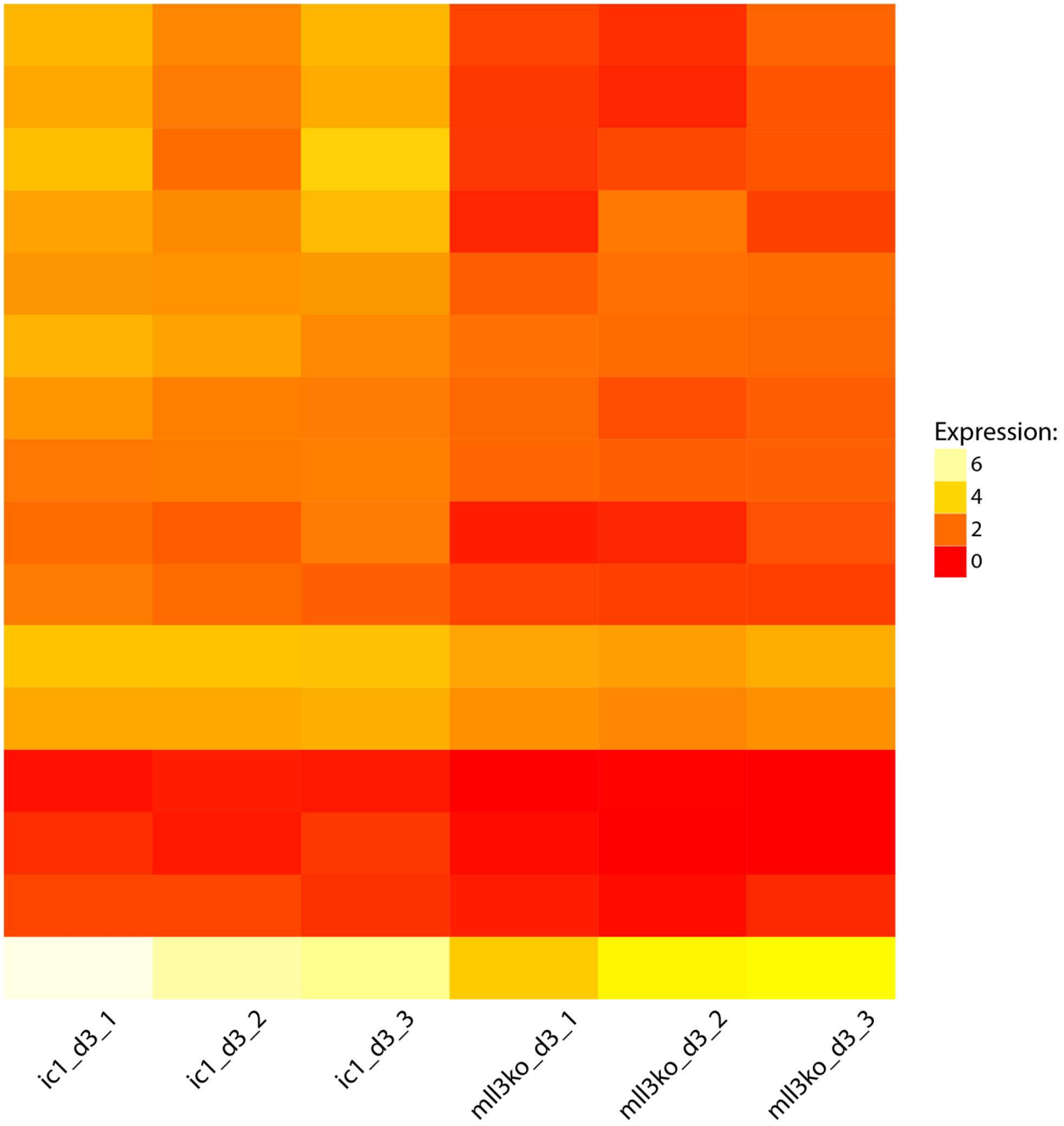


Figure 5.23: Heatmap of day 3 Heart genes

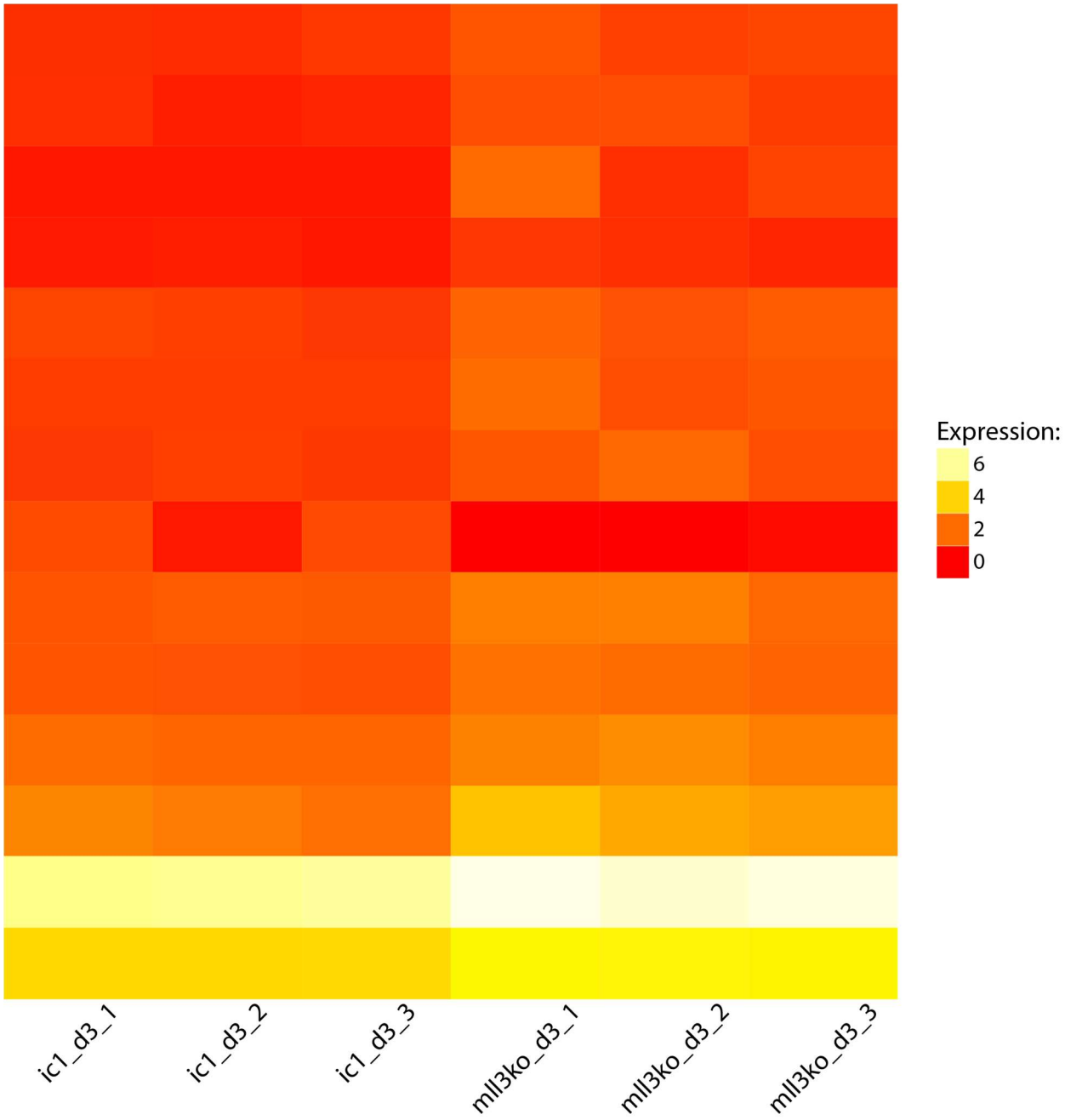


Figure 5.24: Heatmap of day 3 Neuron genes

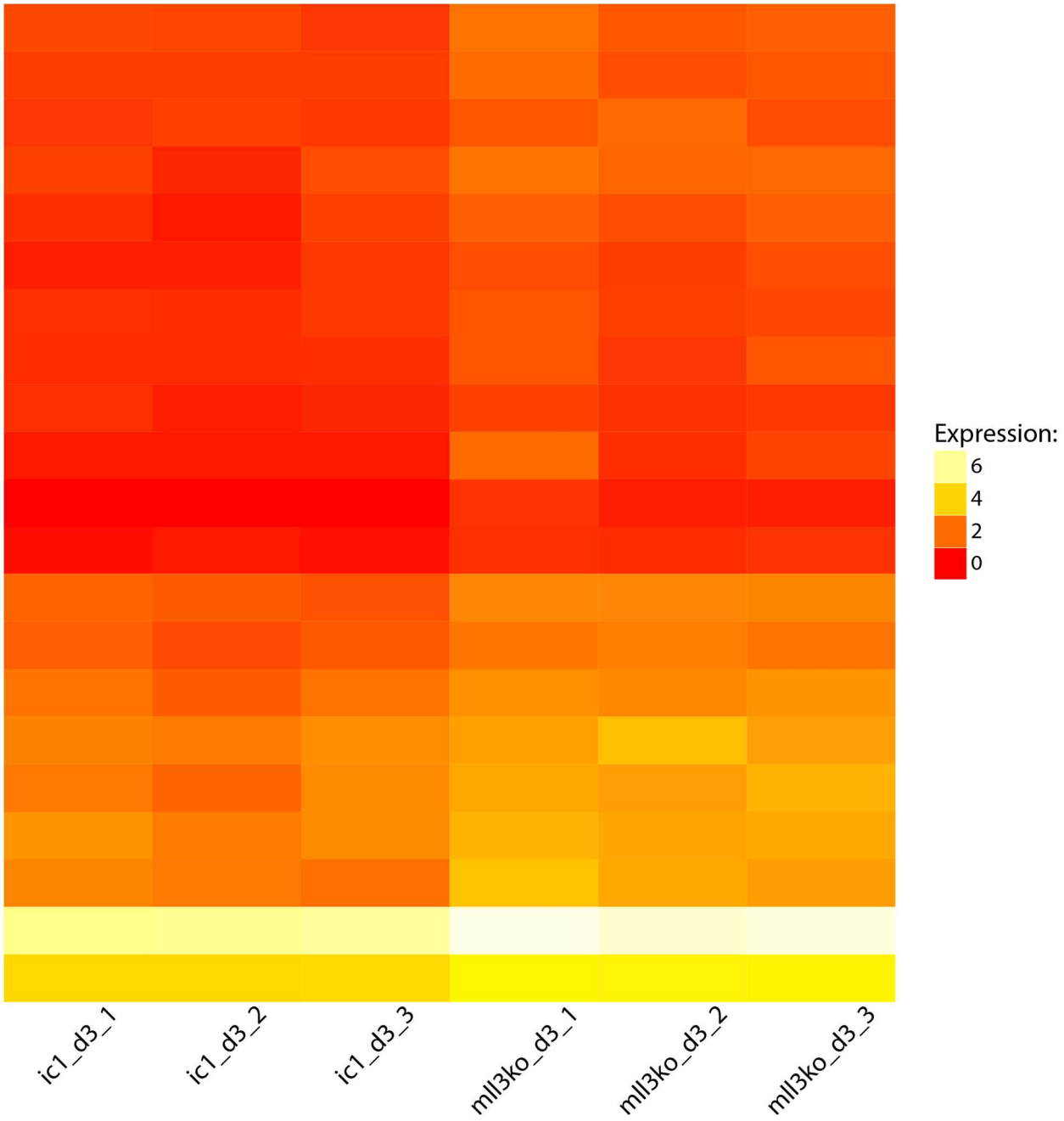


Figure 5.25: Heatmap of day 3 Patterning genes

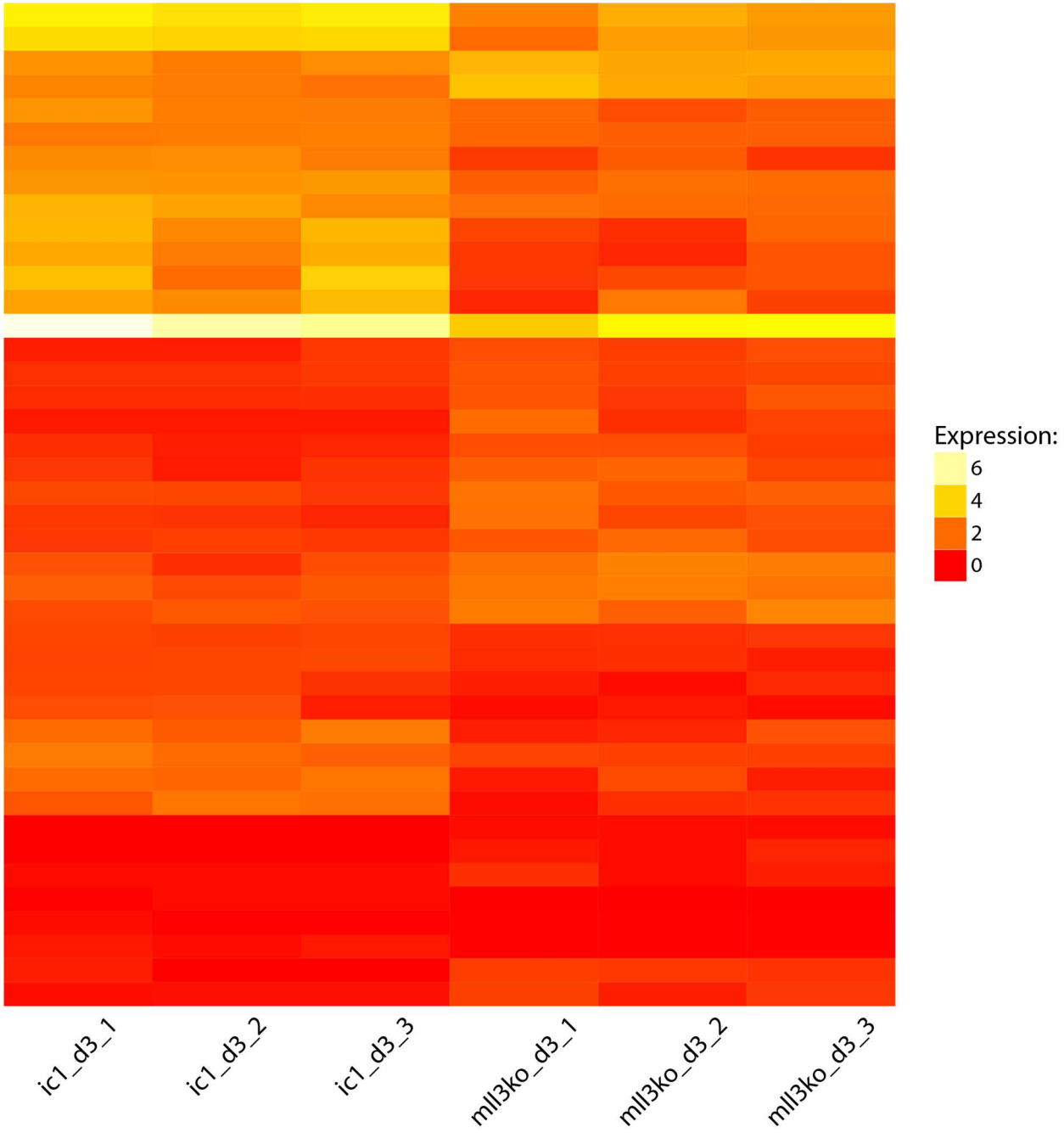


Figure 5.26: Heatmap of day 3 Tissue Specification genes

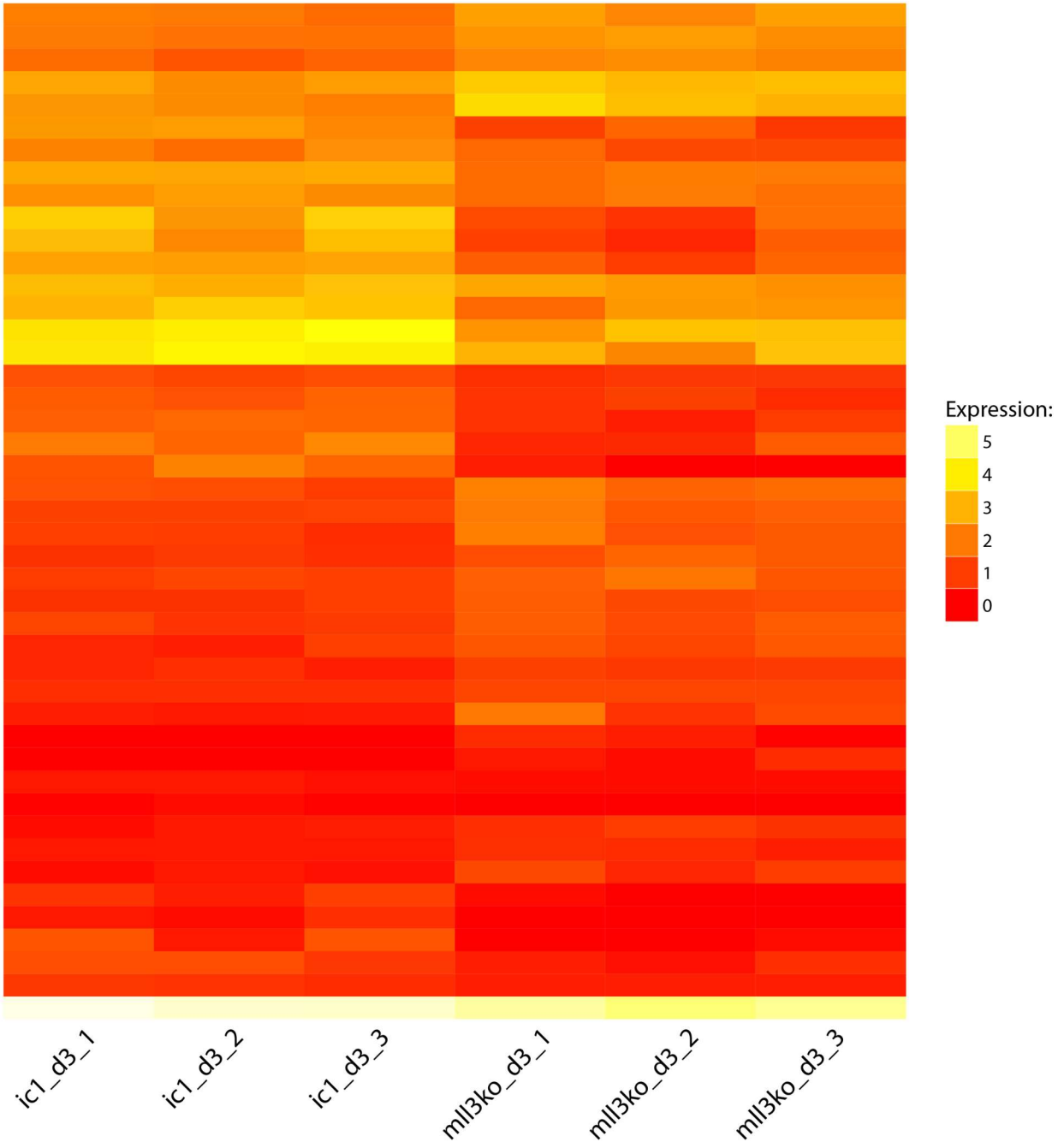


Table 5.1 Genes Significantly upregulated in Day 8 WT cells

A2M	C2orf74	CYTH3	FAM86C1	IFI27L2	MED15	OSGIN2	RAD9A	SLC38A9	TLE4	ZFP3
AACS	C2orf81	D2HGDH	FAM86JP	IFT22	MED16	P2RY2	RARRES3	SLC39A11	TM6SF1	ZFP30
AASDH	C2orf88	DALRD3	FAM92A1	IFT52	MED22	PAAF1	RASA4B	SLC44A3	TMC3	ZFP57
ABCA11P	C3	DBF4B	FAXDC2	IFT74	MEF2B	PABPC1	RASAL3	SLC4A3	TMC6	ZFYVE19
ABCB10	C4A	DBNL	FBF1	IFT80	MEF2BNB	PABPC1L	RASGEF1B	SLC4A8	TMEM120A	ZNF101
ABCB9	C4B_2	DBP	FBXO3	IFT81	MEGF10	PAIP1	RASGRP4	SLC50A1	TMEM141	ZNF132
ABCC4	C5orf63	DCAF11	FBXO36	IK	MEIKIN	PAK1	RASL12	SLC52A2	TMEM144	ZNF135
ABCF1	C9orf64	DCN	FBXO8	IKBKB	METTL12	PAK3	RAVER1	SLC6A13	TMEM164	ZNF138
ABHD16A	CACNA1F	DCP1B	FCER1G	IKZF1	METTL14	PAM	RBAK	SLC7A10	TMEM169	ZNF155
ABHD2	CALCOCO2	DCTN2	FCN3	IL17RC	METTL17	PAPLN	RBL2	SLC7A4	TMEM170A	ZNF177
ABLIM3	CALD1	DDC	FDPS	IL1RAPL2	METTL23	PAPPA2	RBM22	SLC7A8	TMEM184B	ZNF202
ACAD10	CALHM2	DDHD1	FES	IL3	METTL7A	PARD3	RBM25	SLC9A6	TMEM209	ZNF205
ACAN	CALM2	DDHD2	FEZ1	IL32	METTL9	PARP16	RBM46	SLC9A7P1	TMEM241	ZNF207
ACAT1	CALM3	DDI2	FGF10	IMP4	MGAT1	PAX8	RBM7	SLCO2B1	TMEM25	ZNF239
ACBD4	CAPG	DDR1	FGR	INCENP	MGEA5	PCDH1	RBMXL1	SLFN11	TMEM54	ZNF254
ACKR1	CARD8	DDT	FMNL3	ING1	MICAL2	PCDHA10	RCAN3	SLIT2	TMEM62	ZNF311
ACOX1	CARS2	DDTL	FOLR2	INPP4B	MICB	PCDHA11	RDH13	SLITRK6	TMEM92	ZNF322
ACP6	CASP10	DDX19B	FOSL2	INPP5J	MKRN2	PCDHA6	REEP4	SLK	TMIGD2	ZNF343
ACPP	CAST	DDX54	FOXH1	IQCK	MKS1	PCDHB5	RELN	SLU7	TMPRSS11E	ZNF346
ACRBP	CBLB	DDX56	FOXI1	IQSEC2	MLF1	PCDHGB3	REXO4	SLX1B	TNFRSF1B	ZNF350
ACSL6	CBLN2	DECR1	FOXJ3	IRAK1	MLH1	PCNX	RFNG	SMARCD2	TNFRSF9	ZNF382
ACTC1	CBWD1	DENND2D	FOXRED1	IRX6	MLST8	PCTP	RGS1	SMIM11A	TNFSF10	ZNF385D
ACTR1A	CBWD2	DES	FPGT-TNNI3K	IST1	MLX	PDCD10	RGS12	SMIM11B	TNFSF13	ZNF415
ACTR2	CBWD5	DGKB	FTCDNL1	ISYNA1	MME	PDCD6IP	RHAG	SMN1	TNIK	ZNF419
ADAM15	CC2D1A	DHDDS	FUT8	ITGA2B	MMP12	PDE1B	RHBDF2	SMN2	TNIP1	ZNF429
ADAM8	CCDC146	DHFRL1	FXN	ITGAM	MOK	PDE6G	RHBDL3	SMPD4	TNK2	ZNF454
ADCK3	CCDC186	DHPS	FXYD1	ITGB4	MPP1	PDLIM3	RHOF	SMTN	TNNI1	ZNF471
ADCY7	CCDC22	DHRS4	FXYD6	ITSN2	MPZL3	PDXDC1	RHOH	SMTNL2	TNNI3	ZNF506
ADD2	CCDC68	DHRS4L1	FXYD6-FXYD2	IZUMO2	MR1	PECAM1	RIF1	SNRNP25	TNNI3K	ZNF512
ADD3	CCDC69	DHRS4L2	GABRB3	JADE2	MRAP2	PECR	RILP	SNRPE	TNNT1	ZNF528
ADGRE2	CCDC74B	DHX30	GABRG2	JOSD1	MRPL10	PEPD	RINL	SNRPN	TNNT2	ZNF542P
ADGRE5	CCDC77	DHX40	GAD1	JPH4	MRPL2	PEX2	RNASE1	SORBS1	TNR	ZNF544
ADGRG6	CCDC88B	DHX58	GADD45B	KARS	MRPL43	PEX5	RNF114	SOST	TNRC6A	ZNF559
ADHFE1	CCL3L1	DIAPH3	GAL3ST1	KAT5	MRRF	PF4	RNF138	SPAG16	TNS2	ZNF563
ADORA2A	CCL3L3	DISC1	GALC	KAT6A	MRVI1	PFKM	RNF2	SPAST	TNXB	ZNF568
AFAP1	CCR10	DLG1	GALNT9	KAT7	MST1R	PGA3	RNF20	SPATA16	TOM1L1	ZNF570
AGAP3	CD163L1	DLGAP4	GALT	KATNA1	MTA3	PGAP2	RNF213	SPATA20	TOMM40L	ZNF572
AGBL4	CD19	DLGAP5	GANC	KCNAB2	MTHFD1	PGF	RNF222	SPATA7	TOX2	ZNF577
AHCYL2	CD300A	DMC1	GART	KCNG3	MTHFSD	PGLYRP4	RNF25	SPATS2	TP53BP1	ZNF578
AHI1	CD300LG	DMD	GAS8	KCNH2	MYB	PGRMC2	RNF31	SPATS2L	TP73	ZNF585A
AK4	CD33	DMGDH	GATA1	KCNH6	MYBPH	PHB2	RNF34	SPECC1	TPCN1	ZNF585B
AKAP1	CD44	DMKN	GATA4	KCNH7	MYCBPAP	PHLDB1	RNF4	SPHK2	TPT1	ZNF589
AKAP8L	CD53	DNAAF3	GATB	KCNIP2	MYL4	PHYKPL	RNFT1	SPI1	TRAF3IP2	ZNF606
AKIP1	CD6	DNAH14	GATSL3	KCNK17	MYO1F	PI16	RNPEP	SPIDR	TRIM26	ZNF611
AKR1C3	CD68	DNAJC4	GBA	KCNMA1	MYO1G	PICALM	ROBO2	SPINK2	TRIM4	ZNF613
ALAS2	CD82	DNAJC6	GBP5	KDM4C	MYO3A	PIGO	ROBO3	SPN	TRIM34	ZNF619
ALKBH6	CDC25B	DNAL4	GCA	KIAA0141	MYO5A	PIK3AP1	ROCK2	SPOCD1	TRIQK	ZNF626
ALMS1	CDC42EP5	DNASE1L2	GCNT1	KIAA0195	MYO5C	PIK3C3	ROGDI	SPOCK3	TRNAU1AP	ZNF630
ALOX5	CDH11	DOCK11	GDA	KIAA1551	MYO9A	PIK3R3	RPE	SQRDL	TROAP	ZNF662
ALPK2	CDH22	DOK3	GDAP1L1	KIF13A	MYOZ1	PIK3R5	RPL10	SQSTM1	TRPC2	ZNF667
ALS2CR11	CDH3	DONSON	GDAP2	KIF23	N4BP2L1	PIN4	RPL8	SRGN	TRPM3	ZNF671
ALS2CR12	CDK11A	DPM1	GFPT2	KIF5C	NAA30	PIP4K2C	RPP38	SRRT	TRPT1	ZNF674
AMHR2	CDK18	DQX1	GGH	KIRREL2	NAALAD2	PISD	RPS15	SSBP1	TSC22D3	ZNF677
AMZ2	CDK8	DSC3	GH1	KLF1	NAALADL1	PITPNM2	RPS3A	SSBP2	TSGA10IP	ZNF689
ANK3	CDKAL1	DSCAM	GIN1	KLF9	NAB1	PIWIL2	RPS9	SSFA2	TSHZ2	ZNF691
ANKRD18B	CDKL4	DSCR3	GLDN	KLHL33	NAGS	PKP2	RPTOR	SSH3	TSPAN31	ZNF726
ANKRD24	CDKN1B	DSG2	GLRA2	KRI1	NAMPT	PLAU	RPUSD4	SSPN	TSPAN32	ZNF74

ANKRD40	CENPT	DTD1	GLT8D1	L3MBTL4	NAP1L1	PLEK	RRNAD1	SSR1	TSPAN4	ZNF778
ANKRD53	CEP152	DUS2	GLUL	LAMA2	NAP1L6	PLEKHA4	RSPO2	SSR2	TSPAN7	ZNF83
ANKRD55	CEP350	DUSP19	GMFG	LAMC2	NAPB	PLEKHA6	RTEL1	SST	TSPEAR	ZNF880
ANKRD6	CEP95	DYNC1LI1	GNAS	LANCL1	NAT6	PLEKHF1	RUFY1	ST3GAL3	TSPYL5	ZNHIT3
AOAH	CES3	DYX1C1	GOLGA8A	LARP4	NAV2	PLEKHH3	RUFY2	STAMBP	TSSK6	ZNRD1
AP1B1	CFAP44	ECHDC1	GOLGA8N	LAT2	NBEAL1	PLEKHM1	RUNDC3A	STARD3	TTC36	ZSCAN1
AP1G1	CFH	ECHDC3	GOSR2	LDHA	NBPF26	PLSCR5	RUNX1	STAT4	TTC39A	ZSCAN29
AP1G2	CFLAR	ECM1	GP6	LDLRAD4	NCAPG2	PML	RUNX1T1	STAT5A	TTC8	ZSCAN32
APLP1	CFP	EDC4	GPATCH4	LEF1	NDE1	PNKP	RWDD4	STK11IP	TYROBP	ZSWIM7
APLP2	CGN	EEF1A1	GPR108	LEFTY2	NDRG1	PNP	RYR1	STON2	TYW1	ZSWIM8
APOL3	CGRRF1	EEF1B2	GPR132	LEPR	NDRG2	PNPT1	S100A1	STPG1	UBA52	
ARHGAP12	CHD7	EFNA3	GPR156	LETMD1	NDUFA4L2	POLA1	S100A13	STRA6	UBE2D3	
ARHGAP26	CHKA	EFR3A	GPR162	LHX6	NDUFAF1	POLD2	SAMD11	STRADA	UBE2F	
ARHGEF10	CHRNB1	EFTUD2	GPRC5C	LIMCH1	NDUFB4	POLE2	SAMSN1	STRC	UBE2I	
ARHGEF18	CIRBP	EGF	GPSM1	LIMK2	NDUFS7	POLR1D	SAR1B	STX3	UBE2L6	
ARHGEF25	CLASP2	EGFL6	GRAMD3	LIN28A	NEB	POLR2J3	SASH3	STXBP1	UBE2NL	
ARHGEF9	CLDN10	EGFLAM	GREB1L	LIN54	NEIL2	POR	SASS6	STXBP2	UBE2V1	
ARRB2	CLEC10A	EHD1	GRIA3	LIPT1	NEK1	POT1	SATB1	STYXL1	UHRF1BP1L	
ARSA	CLEC2D	EHHADH	GRIP2	LITAF	NELFA	POTEE	SCAPER	SUCNR1	UMPS	
ARSG	CLIC1	EHMT1	GRM4	LMF1	NEMP1	PPARA	SCG5	SUGP2	UNC13D	
ART4	CLK3	EHMT2	GRN	LOC100506403	NEO1	PPARG	SCIN	SULT4A1	UNC45A	
ART5	CLMN	EIF2AK2	GSDMD	LOC101928269	NFASC	PPFIBP1	SCN5A	SUN1	UQCRB	
ASAH1	CLPB	EIF2D	GSN	LOC102724488	NFATC1	PPIP5K2	SCP2	SURF1	UQCRFS1	
ASB1	CLTA	EIF4A1	GSR	LOC102724985	NFE2	PPP1R36	SCRN1	SUV420H1	URAHP	
ASH2L	CLUL1	EIF4A2	GSTM4	LOC105369230	NFKBIL1	PPP2CB	SCRN2	SV2B	USP39	
ASTN1	CLYBL	EIF4E	GSTM5	LOC105369236	NFS1	PPP2R3C	SCYL1	SYNE1	VAMP1	
ATE1	CMBL	ELMOD3	GSTZ1	LOC105369261	NHLRC3	PPP2R5D	SDK1	SYNE2	VAMP7	
ATG16L1	CNOT2	ELP6	GTF2H2C	LOC105369264	NIF3L1	PPP4R1	SEC13	SYNPO2L	VAV1	
ATG3	CNOT8	EMB	GTF2H2C_2	LOC155060	NIN	PPP6R3	SEC14L1	SYT15	VEGFA	
ATL1	CNPY2	EML2	GUF1	LOC440311	NLGN1	PPT1	SEC31A	SYT2	VEZT	
ATL2	COBLL1	EML3	GYPB	LOC641367	NLN	PRCC	SELL	SYT5	VIPR1	
ATM	COG4	EMP3	GYPE	LPAR5	NLRP2	PRIMPOL	SELM	SYVN1	VPS26B	
ATP1A2	COL1A1	ENGASE	HACE1	LPCAT2	NMU	PRKCB	SELPLG	TAB1	VPS28	
ATP2C1	COL1A2	ENO4	HADH	LPL	NOL8	PRKCH	SEMA3E	TAC3	VPS39	
ATP6V0A1	COL5A1	ENOSF1	HAND1	LRRC24	NOSIP	PRKCQ	SEMA4F	TACR1	VRK3	
ATP8B4	COL6A1	ENPP1	HAND2	LRRC37B	NOSTRIN	PRKCSH	SEMA5B	TADA2A	VWDE	
ATXN2	COL6A3	ENPP2	HAP1	LRRC61	NOTCH4	PRKCZ	SEMA6C	TAGLN	WAC	
AXL	COQ5	ENTPD1	HAPLN1	LRRC63	NPNT	PRMT1	SENP2	TAOK3	WASF1	
AZI2	CORO7	ENTPD8	HAS2	LRRIQ3	NQO1	PROK1	SEPP1	TARP	WBP1	
AZIN2	COX6B2	EPB41L2	HBE1	LSM12	NQO2	PRPH	4-Sep	TATDN3	WDR19	
B2M	CPB1	EPOR	HDAC10	LST1	NR1I3	PRPSAP1	SERHL2	TAZ	WDR45	
BAG6	CPED1	EPS15	HDAC11	LTBP4	NRBP2	PRR13	SERPINH1	TBC1D17	WDR74	
BAIAP3	CPNE4	EPS8L1	HDDC3	LTK	NRG1	PRR14	SF3A1	TBC1D26	WDR89	
BBS2	CPNE5	EPS8L3	HDGFRP2	LUC7L2	NRXN1	PRSS57	SFRP5	TBC1D3I	WDR90	
BCAS3	CPS1	ERBB2	HECTD3	LUC7L3	NRXN3	PRTG	SGCA	TBL1XR1	WEE1	
BCAS4	CPSF3L	ERCC2	HERC5	LYPD6B	NSF	PRX	SGSM3	TBX5	WFDC5	
BCLAF1	CRABP2	ERI1	HGF	LYZ	NSMAF	PSMC6	SH2B1	TCAIM	WLS	
BCR	CREB3L4	ERLIN2	HIST1H1A	M1AP	NSMCE2	PSMD6	SH2D3A	TCEANC	WNT9B	
BEND5	CRIP1	ESPNL	HIST1H3I	MACROD2	NSMF	PSME1	SH2D3C	TCEB3	WRNIP1	
BHLHB9	CRTAC1	ESPNP	HIST1H4F	MADD	NT5C	PTAFR	SH3BGR	TCP10L	WWP2	
BMPER	CRYAB	EXOC7	HIST1H4L	MAGEA2	NT5C3A	PTGDS	SHISA5	TCP11	XIRP2	
BMPR1A	CS	EXOSC3	HK1	MAGEA2B	NTRK1	PTGER3	SKAP1	TCTA	XPNPEP1	
BNC2	CSAG3	EZH1	HKR1	MAGEA3	NUCB2	PTGIS	SKIV2L	TDG	XPNPEP2	
BNIP2	CSF1R	F2RL2	HLA-DRB1	MAGEH1	NUDT1	PTGS1	SLAIN2	TEAD4	XPO1	
BORCS8	CSF3R	FAH	HNRNPUL1	MAN2B2	NUP37	PTK7	SLC12A6	TECPR1	XPO6	
BROX	CSTF1	FAM117A	HOXA3	MAP2K5	NUP43	PTN	SLC12A9	TECRL	XRCC3	
BRWD1	CTF1	FAM122C	HOXB2	MAP3K3	NUP93	PTP4A1	SLC13A3	TESK2	XRCC6	
BTBD9	CTGLF12P	FAM131A	HOXB3	MAP3K7CL	NUSAP1	PTPN20	SLC15A4	TET2	XRRA1	
BTK	CTNNA2	FAM156A	HOXB9	MAP4K1	NXF3	PTPN6	SLC16A3	TFDP1	YAE1D1	
BTN3A3	CTNNB1	FAM156B	HPS4	MAP4K4	NXN	PTPRF	SLC16A4	TFPI	YDJC	
BTNL9	CTNNBL1	FAM20A	HSCB	MAPK10	OBFC1	PTPRH	SLC1A5	TGFB1I1	YIPF5	

C11orf21	CTNND2	FAM214B	HSD17B4	MAPK11	OCIAD2	PTPRU	SLC22A5	TGFBI	YWHAE
C11orf57	CTSB	FAM222B	HTATIP2	MAPK4	ODF2	PUS7L	SLC25A13	TGFBR3	YWHAZ
C12orf4	CTSF	FAM227B	HVCN1	MAPKAP1	ODF3B	PWP1	SLC25A30	THAP7	YY1AP1
C12orf60	CUX1	FAM234A	HYAL1	MAST4	OPN3	PYCR1	SLC25A37	THAP9	ZAP70
C16orf45	CUZD1	FAM24B	HYAL3	MATN2	OPN5	PYROXD1	SLC26A10	THEMIS2	ZBED6CL
C19orf25	CXCL10	FAM45A	HYDIN	MAX	OPTN	RAB28	SLC2A8	THOC5	ZBTB10
C19orf57	CXCL5	FAM50A	HYI	MAZ	OR2AG1	RAB37	SLC30A5	THRA	ZBTB7C
C1orf122	CXXC5	FAM65B	IAH1	MCCC2	OR7E47P	RAB3C	SLC32A1	THYN1	ZCCHC10
C1orf228	CYHR1	FAM65C	IBTK	MCFD2	OSBPL3	RAB43	SLC35B1	TIAM2	ZEB2
C1QTNF4	CYP1A1	FAM72A	ICA1	MDGA1	OSBPL6	RABGAP1L	SLC35B2	TIPIN	ZFAND1
C1R	CYP21A2	FAM72C	ICAM1	MDH1	OSBPL9	RAD21	SLC35F2	TJAP1	ZFAND2B
C2orf70	CYP4F30P	FAM72D	IFI16	MECOM	OSCAR	RAD51B	SLC38A6	TLE3	ZFAND6

Table 5.2: Genes significantly upregulated in Day 8 KMT2C KO cells.

A2M	ATAD3A	COL15A1	FLNB	HPGD	LAMB3	NAV3	PHC1	SCD5	SULT1E1
ABCF1	BANP	CREM	FLT4	HSPA12A	LGALS3	NBL1	PILRB	SEL1L3	TAF1C
ACACB	BIRC7	CXCR4	FOXA1	IGFBP3	LOC100506136	NKX2-5	PLEKHG4B	SELE	TCF4
AFP	BTBD11	DAXX	FOXA2	IGSF9	LOC101928143	NOL8	PNMA3	SEMA6B	TF
AGT	C2CD4B	DPEP1	G0S2	IGSF9B	LOC643733	NPAS2	PRICKLE4	SERPINA1	TG
AHSG	C2orf72	DPYS	GEMIN4	IHH	LRWD1	NPR3	PRR26	SHFM1	TIMM22
AMN	C7orf76	DUSP4	GLDC	IL17RD	MAFB	NPTX2	PTCHD1	SHH	TSC2
APCDD1	CDC45	DUSP9	GPC3	ISL1	MAZ	NR4A1	PTPRM	SLC30A1	TTR
APOA1	CDH20	EIF1B	GRB10	ITGA2	MCF2L	NRD1	RAD51B	SLC35F1	VIL1
APOA4	CDK10	ELMO2	GSG1L	ITGA4	MDFIC	NSFL1C	RBM14-RBM4	SLC39A5	VPS53
APOB	CERKL	F2	HAVCR1	KIF19	MELK	NUDT5	RBP4	SLC52A2	WDR47
ARFGEF3	CHST1	FAM57A	HMGCS2	KLK6	MEP1A	NUMB	RNF165	SON	YWHAE
ARVCF	CILP2	FAM64A	HNF4A	KRT7	MGST3	OSBPL3	ROBO4	SOX11	ZACN
ASGR1	CNOT8	FGB	HOXD10	LAMA4	MTTP	PCSK6	RPL22L1	SPTBN5	ZEB1
ASGR2	COL13A1	FGG	HOXD11	LAMB1	MYL3	PDE6B	RPS20	STC1	ZNF512

Table 5.3 Overlapping H3K4me1 and open chromatin sites

Sample	Total ChIP Peaks	Total ATAC-seq Peaks	Overlapping ChIP and ATAC Peaks
WT	16146	17362	12536
KMT2C KO	1214	3277	153

Table 5.4 Significantly Upregulated Genes in Day 3 WT cells

ACP5	CAMK2N2	DPPA4	GDF3	INPP5F	MIB2	POU5F1	SHANK2	U2AF1	ZNF85
ACTB	CAMKV	DPPA5	GDPD2	INPP5J	MIF4GD	PPAP2C	SIPA1	U2AF1L5	ZNF850
ACTRT3	CAV1	ECHDC3	GHRHR	IPO13	MLKL	PPP2R2B	SLC25A29	UBE2D3	ZNF880
ADAM15	CBR3	EEF1A1	GMPR2	IZUMO2	MMP25	PPP2R2C	SLC25A39	UCKL1	ZNHIT3
ADCY2	CCDC141	EFTUD2	GNAS	JADE2	MNS1	PRDM14	SLC29A1	ULK4P1	ZSCAN1
ADNP	CCDC152	EML2	GPS1	JMJD1C	MTHFD1	PRKAR1A	SLC2A14	ULK4P2	
AFF1	CCDC47	EOMES	GRHL2	JUP	MUM1	PSKH2	SLC38A5	UNC13A	
AKAP1	CCM2	EPB41L4B	GRID2	KCNK12	MYBPC2	PTPN14	SLC43A1	UQCC1	
ANK3	CD177	EPHA1	GRM4	KCNK17	MYH14	QRICH2	SLC45A4	UQCRFS1	
ANKLE1	CDH23	EPHX3	GSC	KCTD2	MYO10	RAB17	SLC4A11	USP14	
ANKRD24	CDR2L	EPS8L1	GSTM5	KIAA0930	NANOG	RAB43	SLPI	USP44	
AP4M1	CENPK	ERBB2	GSTO1	KIFC3	NCAN	RABGAP1L	SMPDL3B	VANGL1	
ARHGAP40	CER1	EVPL	GSTO2	KIRREL2	NDE1	RAP1GAP2	SMUG1	VASH2	
ARHGDIA	CERS1	EXOSC5	GYLTL1B	L1TD1	NDRG4	RASSF3	SOX8	VGLL3	
ARMC8	CHGA	FAM124A	HAS3	LASP1	NKAPL	RBFOX3	SPATA13	VRTN	
ARNTL	CHRNA3	FAM155B	HBA1	LGALS7	NLRP2	RBM46	SRY	VSIG10	
ASCL2	CHRNB4	FAM174B	HBA2	LGR5	NLRP7	RDH13	STAU1	WDR59	
ATXN7L1	CMTM3	FAM20A	HBE1	LHX1	NME2	REXO2	STON2	WDR74	
B4GALNT3	CNKSR3	FAM20C	HDAC3	LITAF	NSG1	RFPL2	SYP	WNT3	
BAG6	CNTN1	FAM24B	HDAC6	LOC101929777	NSMCE4A	RHBDF2	SYT3	YIPF1	
BCL11A	COMT	FAT3	HERPUD1	LOC102723996	OAS1	RLN2	TADA2B	ZBED6CL	
BNIP3P9	COX7A1	FBLN1	HES3	LRRC45	OGFOD2	RNF2	TBC1D2	ZDHHC22	
BRINP1	CPNE7	FERMT1	HHLA1	LRRTM1	OTX2	RNF41	TDGF1	ZFP28	
BRSK2	CRABP1	FGF17	HIST1H1A	LZTS1	PAIP2	RNFT2	TDP1	ZIC2	
C10orf2	CRMP1	FGFR1	HIST1H3I	MAD2L2	PAQR4	RPL15	TEAD4	ZIC5	
C19orf33	CRYBB1	FOPNL	HIST1H4F	MAGEA2	PCBP2	RPL3L	TECR	ZIK1	
C1QBP	CS	FOXB1	HIST1H4L	MAGEA2B	PCDH1	RPRML	TGFBR3	ZNF132	
C22orf15	CSMD1	FOXD3	HIST2H4A	MAGEH1	PCDHAC1	RPS17	THEM5	ZNF217	
C2orf81	CST6	FOXK2	HIST3H2A	MALL	PCDHB5	RPS2P32	TJP1	ZNF253	
C2orf88	CTSF	FRS2	HIST3H2BB	MANEAL	PCSK1N	RTN4	TJP2	ZNF350	
C5orf63	CUZD1	FSHR	HKR1	MAP3K9	PDCD6IP	RTP1	TMEM104	ZNF385D	
C6orf136	CYP1B1	FZD5	HNRNPH1	MAP7	PDZD4	RUNX1T1	TMEM132B	ZNF429	
C9orf135	DAZAP1	GABRB3	HNRNPM	MAPK9	PFKFB3	SAC3D1	TNNI3	ZNF506	
C9orf64	DBNDD1	GABRQ	HPCAL1	MARVELD3	PIKFYVE	SCGB3A2	TRIM27	ZNF528	
CA2	DKC1	GALK2	HPDL	MATK	PIWIL2	SEC14L1	TRIML2	ZNF585A	
CACNA1B	DKK4	GAS8	HSPA8	MDGA2	PLEC	SELENBP1	TRNP1	ZNF585B	
CACNG5	DMRTB1	GATA6	ICOSLG	MECR	PLEKHF1	SELT	TRPM2	ZNF667	
CALB1	DNAJC14	GATAD2B	IDO1	MEN1	PLEKHG6	SENP2	TSACC	ZNF677	
CALB2	DND1	GCFC2	IGFBP2	MESP2	POC1B	SEPHS1	TTC9	ZNF74	
CALCR	DNMT3B	GDF1	IGFLR1	MGAT4C	POU2AF1	SFRP2	TTN	ZNF829	

Table 5.5 Significantly Upregulated Genes in Day 3 KMT2C KO cells

ABR	CCM2	DGKH	FOXD1	HN1L	LPIN1	NECAB2	PEMT	SATB2	TM9SF4
ACTA2	CD59	DHRS3	FOXQ1	HOXA4	LRRC4C	NELL2	PGM3	SCD5	TNFRSF19
ADD1	CDH11	DIS3L2	FOXS1	HOXB6	LTBP4	NKX2-5	PHC2	SDC2	TPCN1
ALDH1A2	CDK11A	DNM2	FREM2	HPS1	MAP3K13	NOL3	PIN1	SELENBP1	TPPP
ALX1	CENPBD1P1	DOK6	FSD1	ID1	MAPK15	NPR3	POMT2	SEMA3D	TRAF4
ANGPTL4	CERS3	DUSP15	FTH1	IHH	MCCC2	NRBP2	PPP1R32	SEPT11	TSPAN4
APOA1	CHAD	DUX4L50	G0S2	ILK	MCOLN3	NRGN	PPP2R4	SEPT3	ULBP1
APOPT1	CHD3	EGFLAM	GEMIN4	IP6K2	MDK	NRP1	PPP2R5C	SHISA5	UNC5C
ARL4C	CIRBP	EHMT2	GLMN	KANSL1	MED15	NSUN6	PRICKLE1	SHMT2	URAD
ASB1	CLK1	EIF4E3	GLOD4	KAZALD1	MED24	NTRK2	PRRC2A	SLC35B1	USP14
BFAR	COTL1	EPB41L3	GNAI2	KCNQ2	MEIS2	NTRK3	PRTG	SNRPN	VDAC2
BMI1	CPA2	EPHB4	GOLGA8A	KCP	METTL3	NUDT1	PTCHD4	SPATA7	VEGFA
BRWD1	CPXM1	EXOC5	GPSM1	KDM2A	MFGE8	NUP107	RABL2A	SPPL3	VPS53
C10orf11	CRB2	EXOSC9	GREB1L	KRT16	MLLT3	NUP58	RABL2B	STRA6	VSIG2
CAD	CSF1	FAM120AOS	GSN	KRT16P3	MMP16	NXN	RAD51	TBL1X	WDR6
CARD8	CST3	FAM134A	HAND2	LARP7	MMP28	PAPPA	RBM20	TCF25	YWHAE
CBWD2	CYB5A	FAM57A	HDAC11	LCN15	MNX1	PCED1A	RGPD5	TCF3	ZBTB16
CBWD5	CYP27A1	FAM84A	HDHD2	LEF1	MOV10	PCGF1	RGPD6	THY1	ZDBF2
CBWD6	DCDC2	FBLIM1	HES7	LINGO1	MTRNR2L4	PCSK5	RNMTL1	TIMM22	ZNF503
CCDC144NL	DCTN2	FNTA	HHIPL2	LIPG	NBPF15	PDPR	RPL9	TLX1	ZNF681

6. Discussion

6.1 Progress in Cancer Research and Infant Leukemia

Over the last several decades, dedicated researchers have vastly increased our understanding of the constellation of diseases known as cancer[8,163,164]. We better understand the genetic basis of many cancers, the specific driver mutations present and the clonal heterogeneity present in tumors[17,165–167]. We have developed promising new therapies and treatments[168–170]. However, over that same period, therapy and outcomes for IL have remained largely unchanged[136,171]. Many different treatment protocols for IL have been tested[26–28], but none has been able to improve outcomes in an appreciable way. Similarly, the same approach to tumor-normal sequencing that have provided a wealth of information in adult cancers[17,18,131,172], has failed to meaningfully inform IL[22]. The failure of standard approaches to achieve meaningful progress suggests that, instead of viewing IL as a disease very much like adult leukemia, but occurring in an infant, we should view the disease as a separate entity with unique genetic features and developmental origins. Indeed, a new paradigm for understanding IL needs to be developed.

6.2 The Genetic Basis of IL

In this work, we provide some explanation as to why IL is distinct from many other cancers. Somatic variation is much less present in IL than in other cancers. We hypothesized that leukemia-predisposing germline variants might make up the majority of the "drivers" required for leukemogenesis. There were hints that this might be the case for years. Cases of concordant leukemia in identical twins were well known. They supported the notion that the genetic environment might play a role in the development or maintenance of leukemic cells. Further

evidence for a germline genetic basis for leukemia is seen in the increased risk of ALL between fraternal twins and non-twin siblings[173]. The evidence that childhood cancer, and IL in particular might be more like an inherited complex trait and not a somatic disease was there, but not conclusive. In this work, we provide the strongest evidence to date that germline variation is an important factor in IL and that it might contribute more to disease than do somatic events.

6.3 The Developmental Context of IL

IL develops before birth. Many signaling pathways, cell types and processes are seen exclusively during *in utero* development. These developmental processes lay the foundation for the huge variety of adult tissues and organs, so small perturbations early on might have serious consequences later on. Both of these facts should be considered in IL. Our hPSC modeling is consistent with the leukemic transformation occurring in cells derived from the primitive hematopoietic program. This is almost certainly not the case in adult leukemias. Even if the cell-of-origin in IL is not primitively derived, the early onset of disease suggests that the transformation likely occurs in a more immature cell-type than leukemias occurring later in life. If this is the case, then they would be expected to have distinct behaviors and responsiveness to treatment relative to their definitively derived counterparts. This possibility is much more likely given the burden of deleterious germline variation present in IL patients. If the disease required somatic mutation to occur, then several of these rare events would have to occur very early on, in the same cell and in cancer causing genes or positions. This is incredibly unlikely. However, if the majority of the variation required to develop IL is present in every cell from the time of conception, then the likelihood of receiving the final insult in the transient, immature cells seen only during development is much higher. This idea remains speculative, as we have not

transformed cells in vitro, but it is compelling to think that many of the differences between IL and leukemias later in life was a result of the cell-off-origin. This is further supported by the two studies in which the leukemia that was generated through *in vitro* hematopoietic differentiation was likely primitively derived[151,152]. Further studies on this point are needed.

6.4 Conclusions

This study provides insights into the biology of IL. It supports the hypothesis that germline variation, rather than somatic events, are the relevant genetic events in this disease. It shows that genes found to have significant deleterious variation in IL patients exert, in at least one case, a strong influence on the developing hematopoietic system. Using this system, it also shows that the developmental origins of IL might be different from any other cancer. While these findings are novel and raise many interesting questions, this is an admittedly modest contribution to IL specifically and the larger world of cancer biology generally. It will be fascinating and rewarding to see how this work inspires future research, and this progress will build on and clarify observations made here.

References

1. Hajdu, S. I., Vadmal, M. & Tang, P. A note from history: Landmarks in history of cancer, part 1. *Cancer* **117,** 1097–1102 (2010).

2. Visvader, J. E. Cells of origin in cancer. *Nature* **469,** 314–322 (2011).

3. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 2. *Cancer* **117,** 2811–2820 (2011).

4. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 3. *Cancer* **118,** 1155–1168 (2012).

5. Hajdu, S. I. A note from history: Landmarks in history of cancer, part 4. *Cancer* **118,** 4914–4928 (2012).

6. Hajdu, S. I. & Darvishian, F. A note from history: Landmarks in history of cancer, part 5. *Cancer* **119,** 1450–1466 (2013).

7. Hajdu, S. I. & Vadmal, M. A note from history: Landmarks in history of cancer, part 6. *Cancer* **119,** 4058–4082 (2013).

8. Varmus, H. Fighting Cancers: Continuing Crucial Research on a 'Constellation' of Diseases. *NIHMedline Plus* **7,** 2–3 (2013).

9. Gospodarowicz, M. & O'Sullivan, B. Prognostic factors in cancer. *Semin. Surg. Oncol.* **21,** 13–18 (2003).

10. Rycaj, K. & Tang, D. G. Cell-of-origin of cancer versus cancer stem cells: Assays and interpretations. *Cancer Res.* **75,** 4003–4011 (2015).

11. de Bono, J. S. & Ashworth, A. Translating cancer research into targeted therapeutics. *Nature* **467,** 543–549 (2010).

12. Meric-Bernstam, F. *et al.* Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. *J. Clin. Oncol.* **33,** 2753–2762 (2015).

13. Jordan, E. J. *et al.* Prospective Comprehensive Molecular Characterization of Lung Adenocarcinomas for Efficient Patient Matching to Approved and Emerging Therapies. *Cancer Discov.* CD-16-1337 (2017). doi:10.1158/2159-8290.CD-16-1337

14. Krajewski, K. M., Braschi-amirfarzan, M., Dipiro, P. J., Jagannathan, J. P. & Shinagare, A. B. Molecular Targeted Therapy in Modern Oncology : Imaging Assessment of Treatment Response and Toxicities. **18,** 28–41 (2017).

15. Jeha, S. *et al.* Comparison between pediatric acute myeloid leukemia (AML) and adult AML in VEGF and KDR (VEGF-R2) protein levels. *Leuk. Res.* **26,** 399–402 (2002).

16. Juhl-Christensen, C. *et al.* Genetic and Epigenetic Similarities and Differences between Childhood and Adult AML. *2Pediatric Blood Cancer* **58,** 525–531 (2012).

17. Vogelstein, B. *et al.* Cancer genome landscapes. *Science (80-.).* **339,** 1546–58 (2013).

18. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

19. Janeway, K. A., Place, A. E., Kieran, M. W. & Harris, M. H. Future of clinical genomics in pediatric oncology. *J. Clin. Oncol.* **31,** 1893–1903 (2013).

20. Chen, X., Pappo, A. & Dyer, M. A. Pediatric solid tumor genomics and developmental pliancy. *Oncogene* **34,** 5207–5215 (2015).

21. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).

22. Andersson, A. K. *et al.* The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat. Genet.* **47,** 330–337 (2015).

23. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–308 (2014).

24. Wang, Q. Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes. *Acta Pharmacol. Sin.* **37,** 143–149 (2016).

25. Howlader N, Noone AM, Krapcho M, Miller D, Bishop K, Kosary CL, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, C. K. SEER Cancer Statistics Review, 1975-2014, National Cancer Institute. *https://seer.cancer.gov/csr/1975_2014/,* (2017).

26. Dreyer, Z. E. *et al.* Intensified Chemotherapy Without SCT in Infant ALL: Results From COG P9407 (Cohort 3). *Pediatr. Blood Cancer* **62,** 419–426 (2015).

27. Pieters, R. *et al.* A treatment protocol for infants younger than 1 year with acute lymphoblastic leukaemia (Interfant-99): an observational study and a multicentre randomised trial. *Lancet* **370,** 240–250 (2007).

28. Koh, K. *et al.* Early use of allogeneic hematopoietic stem cell transplantation for infants with MLL gene-rearrangement-positive acute lymphoblastic leukemia. *Leukemia* **29,** 290–296 (2015).

29. Hilden, J. M. *et al.* Analysis of prognostic factors of acute lymphoblastic leukemia in infants : report on CCG 1953 from the Children ' s Oncology Group. *Blood* **108,** 441–451 (2006).

30. Guest, E. M. & Stam, R. W. Updates in the biology and therapy for infant acute lymphoblastic leukemia. *Curr. Opin. Pediatr.* **29,** 1 (2016).

31. Sanjuan-pla, A. *et al.* Revisiting the biology of infant t(4;11)/MLL-AF41 B-cell acute lymphoblastic leukemia. *Blood* **126,** 2676–2686 (2015).

32. Chow, E. J. *et al.* Decreased Adult Height in Survivors of Childhood Acute Lymphoblastic Leukemia: A Report for the Childhood Cancer Survivor Study. *J. Pediatr.* **150,** 370–375 (2007).

33. Silverman, L. B. Acute Lymphoblastic Leukemia in Infancy. *Pediatr. Blood Cancer* **49,** 1070–3 (2007).

34. Pui, C.-H. *et al.* Extended Follow-up of Long-Term Survivors of Childhood Acute Lymphoblastic Leukemia. *N. Engl. J. Med.* **349,** 640–649 (2003).

35. Greaves, M. When one mutation is all it takes. *Cancer Cell* **27,** 433–434 (2015).

36. Montes, R. *et al.* Enforced expression of MLL-AF4 fusion in cord blood CD34+ cells enhances the hematopoietic repopulating cell function and clonogenic potential but is not sufficient to initiate leukemia. *Blood* **117,** 4746–4758 (2011).

37. Bursen, A. *et al.* The AF4·MLL fusion protein is capable of inducing ALL in mice without requirement of MLL·AF4. *Blood* **115,** 3570–3579 (2010).

38. Bueno, C. *et al.* FLT3 activation cooperates with MLL-AF4 fusion protein to abrogate the hematopoietic specification of human ESCs. *Blood* **121,** 3867–3879 (2013).

39. Eden, T. Aetiology of childhood leukaemia. *Cancer Treat. Rev.* **36,** 286–297 (2010).

40. Ross, J. A., Potter, J. D., Reaman, G. H., Pendergrass, T. W. & Robison, L. L. Maternal exposure to potential inhibitors of DNA topoisomerase II and infant leukemia (United States): Report from the Children's cancer group. *Cancer Causes Control* **7,** 581–590 (1996).

41. Wiemels, J. L. *et al.* A lack of a functional NAD(P)H:quinone oxidoreductase allele is selectively associated with pediatric leukemias that have MLL fusions. *Cancer Res.* **59,** 4095–4099 (1999).

42. Wiemels, J. L. *et al.* Methylenetetrahydrofolate reductase (MTHFR) polymorphisms and risk of molecularly defined subtypes of childhood acute leukemia. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 4004–4009 (2001).

43. Gruhn, B. *et al.* Prenatal origin of childhood acute lymphoblastic leukemia, association with birth weight and hyperdiploidy. *Leukemia* **22,** 1692–1697 (2008).

44. Jagannathan-bogdan, M. & Zon, L. I. Hematopoiesis. **2467,** 2463–2467 (2013).

45. Dzierzak, E. & Speck, N. A. Of lineage and legacy: the development of mammalian hematopoietic stem cells. *Nat. Immunol.* **9,** 129–136 (2008).

46. Bloom, W. & Bartelmez, G. W. Hematopoiesis in young human embryos. *Am. J. Anat.* **67,** 21–53 (1940).

47. Palis, J., Robertson, S., Kennedy, M., Wall, C. & Keller, G. Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126,** 5073–5084 (1999).

48. Davidson, A. J. & Zon, L. I. The 'definitive' (and 'primitive') guide to zebrafish hematopoiesis. *Oncogene* **23,** 7233–7246 (2004).

49. Murray, P. D. F. The Development in vitro of the Blood of the Early Chick Embryo. *Proc. R. Soc.* doi:doi:10.1086/303379

50. Choi, K., Kennedy, M., Kazarov, A., Papadimitriou, J. C. & Keller, G. A common precursor for hematopoietic and endothelial cells. *Development* **125,** 725–32 (1998).

51. Fehling, H. J. Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. *Development* **130,** 4217–4227 (2003).

52. Beaupain, D., Martin, C. & Dieterlen-Lievre, F. Are Developmental Hemoglobin Changes Related to the Origin of Stem Cells and Site of Hematopoiesis? *Blood* **53,** 212–225 (1979).

53. Barker, J. E. Development of the mouse hematopoietic system. *Dev. Biol.* **18,** 14–29 (1968).

54. Ditadi, A., Sturgeon, C. M. & Keller, G. A view of human haematopoietic development from the Petri dish. *Nat. Rev. Mol. Cell Biol.* **18,** 56–67 (2016).

55. Palis, J. Hematopoietic stem cell-independent hematopoiesis : emergence of erythroid , megakaryocyte , and myeloid potential in the mammalian embryo. **590,** 3965–3974 (2016).

56. Ditadi, A. *et al.* Human definitive haemogenic endothelium and arterial vascular endothelium represent distinct lineages. *Nat. Cell Biol.* **17,** 580–591 (2015).

57. Sabin, F. R. Studies on the origin of blood-vessels and of red blood-corpuscles as seen in the living blastoderm of chicks during the second day of incubation. 213–262 6 plates. (1920). at <file://catalog.hathitrust.org/Record/001639026>

58. de Bruijn, M. F. T. R. Definitive hematopoietic stem cells first develop within the major arterial regions of the mouse embryo. *EMBO J.* **19,** 2465–2474 (2000).

59. Li, Z. *et al.* Mouse embryonic head as a site for hematopoietic stem cell development. *Cell Stem Cell* **11,** 663–675 (2012).

60. Swiers, G. *et al.* Early dynamic fate changes in haemogenic endothelium characterized at the single-cell level. *Nat. Commun.* **4,** (2013).

61. Boisset, J.-C. *et al.* In vivo imaging of haematopoietic cells emerging from the mouse aortic endothelium. *Nature* **464,** 116–120 (2010).

62. Godin, I. & Cumano, A. The hare and the tortoise: an embryonic haematopoietic race. *Nat. Rev. Immunol.* **2,** 593–604 (2002).

63. Shivdasani, R. A., Mayer, E. L. & Orkin, S. H. Absence of blood formation in mice lacking the T-cell leukaemia oncoprotein tal-1/SCL. *Nature* **373,** 432–434 (1995).

64. Robb, L. *et al.* The scl gene product is required for the generation of all hematopoietic lineages in the adult mouse. *EMBO J.* **15,** 4123–4129 (1996).

65. Robb, L. *et al.* Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 7075–7079 (1995).

66. Porcher, C. *et al.* The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. *Cell* **86,** 47–57 (1996).

67. Yamada, Y. *et al.* The T cell leukemia LIM protein Lmo2 is necessary for adult mouse hematopoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 3890–3895 (1998).

68. Warren, A. J. *et al.* The Oncogenic Cysteine-rich LIM domain protein Rbtn2 is essential for erythroid development. *Cell* **78,** 45–57 (1994).

69. Tsai, F.-Y. *et al.* An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371,** 221–226 (1994).

70. Wang, Q. *et al.* The CBFbeta subunit is essential for CBFalpha2 (AML1) function in vivo. *Cell* **87,** 697–708 (1996).

71. Castilla, L. H. *et al.* Failure of embryonic hematopoiesis and lethal hemorrhages in mouse embryos heterozygous for a knocked-in leukemia gene CBFB-MYH11. *Cell* **87,** 687–696 (1996).

72. Yergeau, D. A. *et al.* Embryonic lethality and impairment of haematopoiesis in mice heterozygous for an AML1-ETO fusion gene. *Nat. Genet.* **15,** 303–306 (1997).

73. North, T. *et al.* Cbfa2 is required for the formation of intra-aortic hematopoietic clusters. *Development* **126,** 2563 LP-2575 (1999).

74. Mucenski, M. L. *et al.* A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis. *Cell* **65,** 677–689 (1991).

75. Wang, Q. *et al.* Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **93,** 3444–9 (1996).

76. Nichogiannopoulou, A., Trevisan, M., Neben, S., Friedrich, C. & Georgopoulos, K. Defects in Hemopoietic Stem Cell Activity in Ikaros Mutant Mice. *J. Exp. Med* **190,** 1201–1213 (1999).

77. Porter, F. D. *et al.* Lhx2 , a LIM homeobox gene , is required for eye , forebrain , and definitive erythrocyte development. *Development* **1242944,** 2935–2944 (1997).

78. Grossmann, M. *et al.* The combined absence of the transcription factors Rel and RelA leads to multiple hemopoietic cell defects. *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11848–53 (1999).

79. Kitajima, K. *et al.* Definitive but not primitive hematopoiesis is impaired in jumonji mutant mice. *Blood* **93,** 87–95 (1999).

80. Scott, E. W. *et al.* PU.1 Functions in a Cell-Autonomous Manner to Control the Differentiation of Multipotential Lymphoid–Myeloid Progenitors. *Immunity* **6,** 437–447 (1997).

81. Creamer, J. P. *et al.* Human definitive hematopoietic specification from pluripotent stem cells is regulated by mesodermal expression of CDX4. *Blood* **129,** blood-2016-11-749382 (2017).

82. Alharbi, R. A., Pettengell, R., Pandha, H. S. & Morgan, R. The role of HOX genes in

normal hematopoiesis and acute leukemia. *Leukemia* **27,** 1000–1008 (2013).

83. Argiropoulos, B. & Humphries, R. K. Hox genes in hematopoiesis and leukemogenesis. *Oncogene* **26,** 6766–6776 (2007).

84. Kennedy, M. *et al.* Development of the hemangioblast defines the onset of hematopoiesis in human ES cell differentiation cultures Plenary paper Development of the hemangioblast defines the onset of hematopoiesis in human ES cell differentiation cultures. **109,** 2679–2687 (2013).

85. Pick, M., Azzola, L., Mossman, A., Stanley, E. G. & Elefanty, A. G. Differentiation of Human Embryonic Stem Cells in Serum-Free Medium Reveals Distinct Roles for Bone Morphogenetic Protein 4, Vascular Endothelial Growth Factor, Stem Cell Factor, and Fibroblast Growth Factor 2 in Hematopoiesis. *Stem Cells* **25,** 2206–2214 (2007).

86. Chadwick, K. *et al.* Cytokines and BMP-4 promote hematopoietic differentiation of human embryonic stem cells. **102,** 906–915 (2003).

87. Wang, Y. & Nakayama, N. WNT and BMP signaling are both required for hematopoietic cell development from human ES cells. *Stem Cell Res.* **3,** 113–125 (2009).

88. Hadland, B. K. *et al.* A requirement for Notch1 distinguishes 2 phases of definitive hematopoiesis during development. *Blood* **104,** 3097–3105 (2004).

89. Vo, L. T. & Daley, G. Q. De novo generation of HSCs from somatic and pluripotent stem cell sources. *Blood* **125,** 2641–2648 (2015).

90. Takayama, N. & Eto, K. in *Platelets and Megakaryocytes: Volume 3, Additional Protocols and Perspectives* (eds. Gibbins, J. M. & Mahaut-Smith, M. P.) 205–217 (Springer New York, 2012). doi:10.1007/978-1-61779-307-3_15

91. Ackermann, M., Liebhaber, S., Klusmann, J. & Lachmann, N. Lost in translation: pluripotent stem cell-derived hematopoiesis. *EMBO Mol. Med.* **7,** 1388–1402 (2015).

92. Ditadi, A. & Sturgeon, C. M. Directed differentiation of definitive hemogenic endothelium and hematopoietic progenitors from human pluripotent stem cells. *Methods* **101,** 65–72 (2016).

93. Kennedy, M. *et al.* T Lymphocyte Potential Marks the Emergence of Definitive Hematopoietic Progenitors in Human Pluripotent Stem Cell Differentiation Cultures. *Cell Rep.* **2,** 1722–1735 (2012).

94. Sturgeon, C. M., Ditadi, A., Awong, G., Kennedy, M. & Keller, G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. *Nat. Biotechnol.* **32,** 554–561 (2014).

95. Ramos, E. *et al.* Population-based rare variant detection via pooled exome or custom hybridization capture with or without individual indexing. *BMC Genomics* **13,** 683 (2012).

96. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

97. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** 164 (2010).

98. Forbes, S. A. *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* 1–26 (2008). doi:10.1002/0471142905.hg1011s57

99. Chang, V. Y., Basso, G., Sakamoto, K. M. & Nelson, S. F. Identification of somatic and germline mutations using whole exome sequencing of congenital acute lymphoblastic leukemia. *BMC Cancer* **13,** 55 (2013).

100. Lupski, J. R., Belmont, J. W., Boerwinkle, E. & Gibbs, R. A. Clan genomics and the complex architecture of human disease. *Cell* **147,** 32–43 (2011).

101. Enciso-Mora, V. *et al.* Common genetic variation contributes significantly to the risk of childhood B-cell precursor acute lymphoblastic leukemia. *Leukemia* **26,** 2212–2215 (2012).

102. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40,** 695–701 (2008).

103. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet.* **4,** (2008).

104. Fearnhead, N. S., Winney, B. & Bodmer, W. F. Rare variant hypothesis for multifactorial inheritance: Susceptibility to colorectal adenomas as a model. *Cell Cycle* **4,** 521–525 (2005).

105. Spector, L. G. *et al.* Maternal diet and infant leukemia: the DNA topoisomerase II inhibitor hypothesis: a report from the children's oncology group. *Cancer Epidemiol Biomarkers Prev* **14,** 651–655 (2005).

106. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).

107. Couto, A. C., Ferreira, J. D., Koifman, S. & Pombo-de-Oliveira, M. S. Familial history of cancer and leukemia in children younger than 2 years of age in Brazil. *Eur. J. Cancer Prev.* **22,** 151–7 (2013).

108. Clarkson, B. D. & Boyse, E. A. Possible Explanatio of the high concordance for actue leukaemia in monozygotic twins. *Lancet* 699–701 (1971).

109. Kadan-Lottick, N. S. *et al.* The Risk of Cancer in Twins: A report from the childhood cancer survivor study. *Pediatr. Blood Cancer* **46,** 476–481 (2006).

110. Van Dijk, B. A., Boomsma, D. I. & De Man, A. J. M. Blood group chimerism in human multiple births is not rare. *Am. J. Med. Genet.* **61,** 264–268 (1996).

111. Quintero, R. *et al.* TWIN-TWIN TRANSFUSION SYNDROME IN A DICHORIONIC-MONOZYGOTIC TWIN PREGNANCY: The End of a Paradigm? *Fetal Pediatr. Pathol.* **29,** 81–88 (2010).

112. Rodriguez, Juan G., Porter, Helen, Sitrrat, Gordon M. , Soothill, P. W. Twin to twin blood

transfusion in a dichorionic pregnancy without the oligohydramnios-polyhydramnios sequence. *Br. J. Obstet. Gynaecol.* **103,** 1049–1056 (1996).

113. French, C. A., Bieber, F. R., Bing, D. H. & Genest, D. R. Twins, placentas, and genetics: Acardiac twinning in a dichorionic, diamniotic, monozygotic twin gestation. *Hum. Pathol.* **29,** 1028–1031 (1998).

114. Foschini, M. P. *et al.* Vascular anastomoses in dichorionic diamniotic-fused placentas. *Int. J. Gynecol. Pathol.* **22,** 359–361 (2003).

115. Gill Super, H. J. *et al.* Clonal, Nonconstitutional Rearrangements of the MLL Gene in Infant Twins With Acute Lymphoblastic Leukemia: In Utero Rearrangement of 11q23. *Blood* **83,** 641–644 (1994).

116. Chuk, M. K., McIntyre, E., Small, D. & Brown, P. Discordance of MLL-rearranged (MLL-R) infant acute lymphoblastic leukemia in monozygotic twins with spontaneous clearance of preleukemic clone in unaffected twin. *Blood* **113,** 6691–6694 (2009).

117. Ruault, M., Brun, M. E., Ventura, M., Roizès, G. & De Sario, A. MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently deleted in myeloid leukaemia. *Gene* **284,** 73–81 (2002).

118. Tenney, K. & Shilatifard, A. A COMPASS in the voyage of defining the role of trithorax/MLL-containing complexes: Linking leukemogensis to covalent modifications of chromatin. *J. Cell. Biochem.* **95,** 429–436 (2005).

119. Ansari, K. I. & Mandal, S. S. Mixed lineage leukemia: Roles in gene expression, hormone signaling and mRNA processing. *FEBS J.* **277,** 1790–1804 (2010).

120. Balakrishnan, A. *et al.* Novel somatic and germline mutations in cancer candidate genes in glioblastoma, melanoma, and pancreatic carcinoma. *Cancer Res.* **67,** 3545–3550 (2007).

121. Dolnik, A. *et al.* Commonly altered genomic regions in acute myeloid leukemia are enriched for somatic mutations involved in chromatin remodeling and splicing. *Blood* **120,** 83–93 (2013).

122. Li, W.-D. *et al.* Exome sequencing identifies an MLL3 gene germ line mutation in a pedigree of colorectal cancer and acute myeloid leukemia. *Blood* **121,** 1478–1479 (2013).

123. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).

124. Machado, C. & Andrew, D. J. D-Titin: A giant protein with dual roles in chromosomes and muscles. *J. Cell Biol.* **151,** 639–651 (2000).

125. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

126. Patel, Z. H. *et al.* The struggle to find reliable results in exome sequencing data: Filtering out Mendelian errors. *Front. Genet.* **5,** 1–13 (2014).

127. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74

(2015).

128. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

129. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP). at <http://evs.gs.washington.edu/EVS/>

130. Hodgkinson, A. & Eyre-Walker, A. Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12,** 756–766 (2011).

131. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).

132. Hu, D. *et al.* The MLL3/MLL4 Branches of the COMPASS Family Function as Major Histone H3K4 Monomethylases at Enhancers. *Mol. Cell. Biol.* **33,** 4745–4754 (2013).

133. Bodian, D. L. *et al.* Germline variation in cancer-susceptibility genes in a healthy, ancestrally diverse cohort: Implications for individual genome sequencing. *PLoS One* **9,** (2014).

134. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).

135. Döhner, H. & Gaidzik, V. I. Impact of genetic features on treatment decisions in AML. *Hematology* **2011,** 36–42 (2011).

136. Brown, P. Treatment of infant leukemias: challenge and promise. *Hematology* 596–600 (2013). doi:10.1182/asheducation-2013.1.596

137. Winters, A. C. & Bernt, K. M. MLL-Rearranged Leukemias—An Update on Science and Clinical Approaches. *Front. Pediatr.* **5,** 11–13 (2017).

138. Hilden, J. M. *et al.* Analysis of prognostic factors of acute lymphoblastic leukemia in infants: report on CCG 1953 from the Children's Oncology Group. *Blood* **108,** 441–51 (2006).

139. Seita, J. & Weissman, I. L. Hematopoietic stem cell: Self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2,** 640–653 (2010).

140. Zhang, Y., Chen, A., Yan, X. M. & Huang, G. Disordered epigenetic regulation in MLL-related leukemia. *Int. J. Hematol.* **96,** 428–437 (2012).

141. Gole, B. & Wiesmüller, L. Leukemogenic rearrangements at the mixed lineage leukemia gene (MLL)-multiple rather than a single mechanism. *Front. Cell Dev. Biol.* **3,** 41 (2015).

142. Ditadi, A. & Sturgeon, C. M. Directed differentiation of definitive hemogenic endothelium and hematopoietic progenitors from human pluripotent stem cells. *Methods* **101,** 65–72 (2016).

143. Creamer, J. P. *et al.* Human definitive hematopoietic specification from pluripotent stem cells is regulated by mesodermal expression of CDX4. *Blood* **129,** blood-2016-11-749382 (2017).

144. Sturgeon, C. M., Ditadi, A., Awong, G., Kennedy, M. & Keller, G. Wnt signaling controls the specification of definitive and primitive hematopoiesis from human pluripotent stem cells. *Nat. Biotechnol.* **32,** 554–561 (2014).

145. Tiyaboonchai, A. *et al.* Utilization of the AAVS1 safe harbor locus for hematopoietic specific transgene expression and gene knockdown in human ES cells. *Stem Cell Res.* **12,** 630–637 (2014).

146. Tenen, D. G. Disruption of differentiation in human cancer: AML shows the way. *Nat. Rev. Cancer* **3,** 89–101 (2003).

147. Yan, M. & Liu, Q. Differentiation therapy: A promising strategy for cancer treatment. *Chin. J. Cancer* **35,** 10–12 (2016).

148. Pirozzi, C., Reitman, Z. & Yan, H. Releasing the Block: Setting Differentiation Free with Mutant IDH Inhibitors. *Cancer Cell* **23,** 570–572 (2013).

149. Tamai, H. *et al.* Activated K-Ras protein accelerates human MLL/AF4-induced leukemo-lymphomogenicity in a transgenic mouse model. *Leukemia* **25,** 888–891 (2011).

150. Stam, R. W. *et al.* Gene expression profiling − based dissection of MLL translocated and MLL germline acute lymphoblastic leukemia in infants Gene expression profiling – based dissection of MLL translocated and MLL germline acute lymphoblastic leukemia in infants. *Blood* **115,** 2835–2844 (2014).

151. Chao, M. P. *et al.* Human AML-iPSCs Reacquire Leukemic Properties after Differentiation and Model Clonal Variation of Disease. *Cell Stem Cell* **20,** 329–344.e7 (2017).

152. Kotini, A. G. *et al.* Stage-Specific Human Induced Pluripotent Stem Cells Map the Progression of Myeloid Transformation to Transplantable Leukemia. *Cell Stem Cell* **20,** 315–328.e7 (2017).

153. Fehling, H. J. Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. *Development* **130,** 4217–4227 (2003).

154. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).

155. Pimentel, H. J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *Nat. Publ. Gr.* (2016). doi:10.1101/058164

156. Dobin, A., Gingeras, T. R. & Spring, C. Mapping RNA-seq Reads with STAR. *Curr. Protoc. Bioinforma.* **51,** (2016).

157. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

158. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

159. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human

genome. ... *Site/Anshulkundaje/Projects/Blacklists (Last Accessed 30 ...* (2013). at <ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byFreeze/jan2011/blacklists/hg19-blacklist-README.pdf%5Cnhttps://sites.google.com/site/anshulkundaje/projects/blacklists>

160. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).

161. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38,** 576–589 (2010).

162. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat. Methods* **12,** 963–965 (2015).

163. Burstein, H. J. *et al.* Clinical Cancer Advances 2017: Annual Report on Progress Against Cancer From the American Society of Clinical Oncology. *J. Clin. Oncol.* **35,** 1341–1367 (2017).

164. Sudhakar, a. History of Cancer, Ancient and Modern Treatment Methods Akulapalli. *J Cancer Sci Ther.* **1,** 1–4 (2010).

165. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* **168,** 613–628 (2017).

166. Punt, C. J. A., Koopman, M. & Vermeulen, L. From tumour heterogeneity to advances in precision treatment of colorectal cancer. *Nat. Rev. Clin. Oncol.* **14,** 235–246 (2016).

167. Prasetyanti, P. R. & Medema, J. P. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol. Cancer* **16,** 41 (2017).

168. Guth, A. M. & Dow, S. Cancer Immunotherapy. *Withrow MacEwen's Small Anim. Clin. Oncol.* **342,** 198–214 (2013).

169. Collins, I. & Workman, P. New approaches to molecular cancer therapeutics. *Nat. Chem. Biol.* **2,** 689–700 (2006).

170. Liu, X. *et al.* Affinity-tuned ErbB2 or EGFR chimeric antigen receptor T cells exhibit an increased therapeutic index against tumors in mice. *Cancer Res.* **75,** 3596–3607 (2015).

171. Stam, R. W., Den Boer, M. L. & Pieters, R. Towards targeted therapy for infant acute lymphoblastic leukaemia. *Br. J. Haematol.* **132,** 539–551 (2006).

172. Dong, H. & Wang, S. Exploring the cancer genome in the era of next-generation sequencing. *Front. Med.* **6,** 48–55 (2012).

173. Hemminki, K. & Jiang, Y. Risks among siblings and twins for childhood acute lymphoid leukaemia: results from the Swedish Family-Cancer Database. *Leukemia* **16,** 297–298 (2002).